

GMWB Riders in a Binomial Framework - Pricing, Hedging, and Diversification of Mortality Risk*

Cody HYNDMAN^{†‡} and Menachem WENGER^{§¶}

July 7, 2016

Abstract

We construct a binomial model for a guaranteed minimum withdrawal benefit (GMWB) rider to a variable annuity (VA) under optimal policyholder behaviour. The binomial model results in explicitly formulated perfect hedging strategies funded using only periodic fee income. We consider the separate perspectives of the insurer and policyholder and introduce a unifying relationship. Decompositions of the VA and GMWB contract into term-certain payments and options representing the guarantee and early surrender features are extended to the binomial framework. We incorporate an approximation algorithm for Asian options that significantly improves efficiency of the binomial model while retaining accuracy. Several numerical examples are provided which illustrate both the accuracy and the tractability of the binomial model. We extend the binomial model to include policy holder mortality and death benefits. Pricing, hedging, and the decompositions of the contract are extended to incorporate mortality risk. We prove limiting results for the hedging strategies and demonstrate mortality risk diversification. Numerical examples are provided which illustrate the effectiveness of hedging and the diversification of mortality risk under capacity constraints with finite pools.

Keywords: variable annuity; GMWB; optimal stopping; hedging; binomial models; mortality

Mathematics Subject Classification (2010): Primary: 91G20; 91G60; Secondary: 91B30; 60G40

JEL Classification: G22, G12, G13, C61, C63

*This paper combines a previous version titled “Pricing and Hedging GMWB Riders in a Binomial Framework” (*arXiv:1410.7453v1*) and the working paper titled “Diversification of mortality risk in GMWB rider pricing and hedging”

[†]Corresponding Author: email: cody.hyndman@concordia.ca

[‡]Department of Mathematics and Statistics, Concordia University, 1455 Boulevard de Maisonneuve Ouest, Montréal, Québec, Canada H3G 1M8.

[§]The Guardian Life Insurance Company of America, New York, NY

[¶]The views and opinions expressed in this paper are those of the individual author(s) and do not necessarily reflect the views of The Guardian Life Insurance Company of America.

1 Introduction

The variable annuity (VA) with guaranteed minimum withdrawal benefit (GMWB) rider was introduced in 2002. These contracts allow for an accumulation period where an initial premium deposited with the insurer is invested in a portfolio of funds selected by the policyholder. The account value (AV) benefits from gains made by the portfolio and a periodic fee is deducted by the insurer. The policy holder can take periodic withdrawals from the AV, up to certain limits, and cumulative withdrawals are guaranteed to return the initial premium over the term of the contract. The contract may be surrendered early, enabling the policyholder to benefit from strong portfolio performance, subject to contingent deferred sales charges (CDSC). At the end of the term, provided the contract has not already been surrendered, the contract may be annuitized for either a fixed term or the remaining life of the policyholder. A large literature on the modeling and pricing of these contracts, as well as other forms of guarantees, has emerged since their introduction to the marketplace. A brief overview of the history of GMWB and similar products as well as the various modeling and pricing approaches can be found in Hyndman and Wenger [15].

Around the time of the financial crisis in 2008 reinsurers stopped offering coverage altogether on GMWB and related guaranteed lifetime withdrawal benefit (GLWB) riders at which point the importance of internal dynamic hedging programs rose rapidly. With this in mind, we consider the problems of pricing and hedging the GMWB product in a discrete time framework consistent with the no-arbitrage principle from financial economics. We propose a binomial asset pricing model for GMWBs assuming optimal policyholder behaviour and construct explicit hedging strategies. An overview of other approaches to policyholder behaviour can be found in Kling et al. [16], Li and Szimayer [17], and the references therein.

The binomial model has several advantages which we believe justify its use in theory and practice. It is significantly simpler to obtain numerical results using the binomial model than many of the approaches which have previously been applied. Under an appropriate parameterization the binomial model converges to the Black and Scholes [3] model, which has been used as the basis for modeling these contracts by a number of authors, and yields good approximations for more complex financial options which lack analytic solutions in the corresponding continuous time pricing models. Through dynamic programming and backward induction algorithms, binomial pricing models can easily be implemented. Further, the binomial model can be calibrated to a volatility surface.

In contrast to Monte-Carlo simulation methods, the binomial approach is well-suited for American-style options with early exercise capability. More importantly an explicit exact hedging strategy can be formulated and implemented. Although binomial methods can be seen as a special case of finite difference methods there are fundamental differences between the two general methods and a thorough comparison of binomial and finite difference methods is provided in Geske and Shastri [12].

Binomial models are ideally suited for non path-dependent products. In such a setting, aside from enabling a simple theoretical framework, it is computationally efficient to obtain reliable numerical results. The GMWB product is path-dependent and we discuss the implications of this and address them by employing an approximation technique. Although in theory the results should converge to those of the continuous withdrawal model where the investment fund is log-normally distributed; due to the non-recombining nature of the account value the suggested method is found to be numerically expensive. We substantially improve the numerical efficiency without sacrificing significant accuracy of results by adopting an approximation method based on Costabile et al. [7].

A binomial valuation approach has previously been considered by Bacinello [1] to price equity-linked life insurance with recurring premiums in the presence of early surrenders. Although the underlying methodology is similar, we deal with the unique features and challenges of modeling GMWB riders for variable annuities. In addition to surrender and mortality, both elements considered by Bacinello [1], we have an endogenously determined trigger date. The nature of the fees and withdrawals further differentiate our work. Whereas Bacinello [1] deals exclusively with pricing, we pay equal attention to the hedging constructions in a binomial model, which is facilitated by the consideration of the perspectives of both the insurer and insured. By focusing on a single product we have the liberty to consider a top-down approach which provides more insight than generic formulations of backward induction schemes.

The remainder of this paper is organized as follows. In Section 2 we present the binomial asset pricing model for variable annuities with GMWBs riders in a restricted model which accounts for equity risk only. We extend the model in Section 3 to allow for surrenders - that is, we incorporate

behaviour risk. In Section 4 we discuss computational considerations in the implementation of the binomial model and we present a binomial approximation algorithm designed to improve numerical efficiency. In Section 5 numerical results using the binomial model are obtained and compared with results from the literature. In Section 6 we extend the binomial model to include mortality risk and death benefits as well as proving mortality diversification results and considering numerical experiments reflecting capacity constraints and finite pools of policy holders. Section 7 concludes and an Appendix contains technical results and proofs.

2 GMWBs in a Binomial Asset Pricing Model

In this section we define and construct a binomial asset pricing model for the variable annuity with GMWB rider. We first introduce the product specifications and notation following Shreve [23] and Duffie [10] for the binomial model and Hyndman and Wenger [15] for the variable annuity with GMWB rider.

2.1 Contract Specifications and Model Framework

At time $t = 0$, a policy, consisting of an underlying variable annuity (VA) contract and a GMWB rider, is issued to a policyholder of age x and an initial premium P is received. We assume no subsequent premiums are paid after time zero. The premium is invested into a fund which tracks the price of a risky asset $S = \{S_t : t \geq 0\}$ with no basis risk. The rider fee rate α is periodically discounted from the account value $W = \{W_t; t \geq 0\}$ as long as the contract is in force and the account value is positive.

The GMWB rider contract specifies a guaranteed maximal withdrawal rate g so that $G := gP$ can be withdrawn annually until the initial premium is recovered regardless of the evolution of $\{W_t\}$. If the account value falls to zero the policyholder continues to make withdrawals at rate G until the initial premium has been recovered. Policyholders may withdraw any amount from the account value not exceeding the remaining account value. However, if annual withdrawals exceed G while the account value is still positive then a surrender charge is applied to the withdrawals and a reset feature may reduce the guarantee value, i.e. the remaining portion of the initial premium not yet recovered. Policyholders also have the option of surrendering early and receiving the account value less a surrender charge. The terminology of lapses and surrenders are used interchangeably. Any guarantee value is forfeited by surrendering.

Assuming a static withdrawal strategy where G is withdrawn annually we set the maturity $T := 1/g$ since the sum of all withdrawals at T is P . At time T the rider guarantee is worthless and the policyholder receives a terminal payoff of the remaining account value if it is positive. This assumption translates over to a real-world trend of no annuitizations and is justified since a high proportion of VAs are not ultimately annuitized.

Consider a financial market consisting of one risky asset S and a riskless money market account growing at a constant continuously compounded risk-free interest rate r . Let n be the number of time-steps per year, $N = T \times n$ the total number of time-steps modelled, and $\delta t = 1/n$ the length of each time-step. For $i \in \mathcal{I}_N^+ := \{1, \dots, N-1, N\}$, write S_i for the asset value at time $i\delta t$. We assume that the insurer can borrow and lend at rate r . Given S_{i-1} , the asset value S_i takes one of two values: $S_{i-1}u$ or $S_{i-1}d$, where u (d) represents an up-movement (down-movement) in the asset value. To rule out arbitrage opportunities and the trivial case of no randomness, u and d must satisfy

$$0 < d < e^{r\delta t} < u \quad (1)$$

as in Shreve [23].

Consider a sequence of N coin tosses. Let $\Omega := \{H, T\}^N$ and $\mathcal{F} := 2^\Omega$. Denote a sample point of Ω by $\bar{\omega}_N := \omega_1 \dots \omega_N := (\omega_1, \dots, \omega_N)$. Consider the stochastic process $\xi = (\xi_i)_{1 \leq i \leq N}$, where $\xi_i : \Omega \mapsto \{u, d\}$ is

$$\xi_i(\bar{\omega}_N) = \begin{cases} u & \text{if } \omega_i = H, \\ d & \text{if } \omega_i = T. \end{cases}$$

Then for any fixed $\bar{\omega}_N$, $\xi_i(\bar{\omega}_N)$ maps i to the growth factor of S in period i . The natural filtration is $\mathcal{F}_i = \sigma(\xi_j; j \leq i)$.

The \mathbb{F} -adapted process $\{S_i\}$ can be expressed as $S_i = S_0 \times \prod_{j=1}^i \xi_j$ where S_0 is the initial value of the risky asset. We write $\bar{\omega}_i = \omega_1 \dots \omega_i$ to refer to the specific path evolution up to time i . For any $j \leq i$, we write

$$\xi_j(\bar{\omega}_i) = \begin{cases} u & \text{if } \omega_j = H, \\ d & \text{if } \omega_j = T. \end{cases}$$

Finally, we replace H and T with u and d respectively when defining Ω , therefore the sample path $\bar{\omega}_N$ refers directly to the evolution of the underlying asset S where each $\omega_j \in \{u, d\}$. Then for any $\bar{\omega}_i$,

$$S_i = S_0 \prod_{j=1}^i \xi_j(\bar{\omega}_i) = S_0 \prod_{j=1}^i \omega_j = S_0 u^{\{\# \text{ of } u \text{ in } \bar{\omega}_i\}} d^{\{\# \text{ of } d \text{ in } \bar{\omega}_i\}}. \quad (2)$$

Beginning with $S_0 = P$, the binomial tree for $\{S_i\}$ is constructed forward in time. For $i \in \mathcal{I}_N^+$, set

$$S_i = \xi_i S_{i-1}.$$

The unique risk-neutral measure \mathbb{Q} is defined on (Ω, \mathcal{F}_N) by

$$\mathbb{Q}(A) := \sum_{\bar{\omega}_N \in A} p^{\{\# \text{ of } u \text{ in } \bar{\omega}_N\}} q^{N - \{\# \text{ of } u \text{ in } \bar{\omega}_N\}}$$

for any set $A \in \mathcal{F}_N$ where

$$p := \frac{e^{r\delta t} - d}{u - d}, \quad (3)$$

is the risk-neutral probability of observing an H at any particular coin toss (observing a u at any particular time step), $q := 1 - p$, and $p > 0$. The constructed probability space is $(\Omega, \mathcal{F}_N, \mathbb{P} = \{\mathcal{F}_i\}_{0 \leq i \leq N}, \mathbb{Q})$. Note that $p \in (0, 1)$ by (1) and there are no $(\mathbb{Q}, \mathcal{F}_N)$ -negligible sets and so all results hold for all $\omega \in \Omega$.

We follow the Cox, Ross, and Rubinstein (CRR) parametrization and set $u = \exp(\sigma\sqrt{\delta t})$ and $d = \exp(-\sigma\sqrt{\delta t})$, where σ is the variance of the continuously compounded rate of return of S . The CRR parametrization leads to the following result.

Proposition 1. *Suppose S_t follows the dynamics given by*

$$dS_t = rS_t dt + \sigma S_t dB_t \quad (4)$$

where B_t is a standard Brownian motion. Consider the binomial model for S_t^n with n time-steps per year under the CRR parametrization. Then for all $t \in [0, T]$, as $n \rightarrow \infty$, S_{nt}^n converges in distribution to S_t , where nt is an integer.

Proof. See Cox et al. [8] or Shreve [24, Exercise 3.8]. \square

Note that the real world probability measure \mathbb{P} is defined similarly but with $\tilde{p} = \frac{1}{2} + \frac{1}{2} \frac{\hat{\mu}}{\hat{\sigma}} \sqrt{\delta t}$ where $\hat{\mu}$ and $\hat{\sigma}$ are the respective empirical mean and variance of the continuously compounded rate of return of S . Under this parametrization, the mean and variance under the binomial model converge to the empirical values in the limit (see Cox et al. [8]).

We specify the underlying assumptions on the variable annuity with GMWB rider that are employed throughout this section.

Assumption 2. *Early surrenders are not allowed. Under the static withdrawal strategy the policyholder receives $G = gP\delta t$ each time period. We set $T := 1/g$. At the end of each period the pro-rated rider fee is first deducted and then the periodic withdrawal is subtracted. We restrict $r > 0$ and denote $\bar{r} := r\delta t$ and $\bar{\alpha} := \alpha\delta t$.*

Remark 1. *We assume T is an integer. Otherwise, the results can be adapted to incorporate the final fractional period. Set $N = \lfloor T \cdot n \rfloor + 1$ and the final period has time length of $T - (\frac{N-1}{n})$ years. All the parameters need to be scaled for the terminal period to reflect the shortened duration.*

Next, we define another binomial tree for the account value W which contains two values at each node. The first component, denoted W_{i-} , is the account value after adjusting for market movements but before fees are deducted or withdrawals are made. The second component, denoted W_i , is the account value after adjusting for fees and withdrawals. We have

$$\begin{aligned} W_0 &= P, \\ W_{i-} &= \frac{S_i}{S_{i-1}} W_{i-1} = \xi_i W_{i-1}, \\ W_i &= \max \{ e^{-\bar{\alpha}} W_{i-} - G, 0 \}, \end{aligned}$$

for $i \in \mathcal{I}_N^+$. Although the tree for the underlying asset $\{S_i\}$ is recombining, the tree for the account value $\{W_i\}$ is non-recombining. For any i there are $i+1$ nodes for S_i but 2^i nodes for W_i on the respective trees. The subtraction of the periodic withdrawals imposes a path dependency on the model.

2.2 Valuation perspectives and decompositions

There are two separate perspectives for valuing the variable annuity with GMWB rider. The first, corresponding to the policyholder, treats the variable annuity and GMWB rider together and values the total payments received over the life of the contract. The second, corresponding to the insurer, considers the embedded optionality of the GMWB rider separately to price and hedge the additional risk. This approach was used in Peng et al. [22] and Hyndman and Wenger [15] in a continuous time setting. In the discrete-time binomial model we obtain similar theoretical results as in the continuous-time setting of Hyndman and Wenger [15]. However, in the discrete-time binomial model we also provide an explicit computational framework for pricing and hedging.

2.2.1 Policyholder valuation perspective

Denote by V_n the value to the policyholder at time n of the remaining payments to be received from the complete contract (VA plus GMWB rider). By the risk neutral pricing formula we obtain the following backward in time recursive relationship

$$\begin{aligned} V_N &= W_N, \\ V_i &= E_{\mathbb{Q}} \left[\sum_{m=i+1}^N G e^{-\bar{r}(m-i)} + e^{-\bar{r}(N-i)} W_N \mid \mathcal{F}_i \right] \\ &= G a_{\overline{N-i}} + e^{-\bar{r}(N-i)} E_{\mathbb{Q}}[W_N \mid \mathcal{F}_i] \end{aligned} \tag{5}$$

for $i \in \mathcal{I}_{N-1}$, with $\mathcal{I}_{N-1} := [0, 1, \dots, N-1]$ and $a_{\overline{m}} = (1 - e^{-\bar{r}m}) / (e^{\bar{r}} - 1)$. For $i = 0$ this reduces to

$$V_0 = G a_{\overline{N}} + e^{-\bar{r}N} E_{\mathbb{Q}}[W_N]. \tag{6}$$

Note that equations (5) and (6) are the discrete-time analogues to the policyholder's valuation given, respectively, by equations (7) and (6) of Hyndman and Wenger [15]. We write $V_0 := V_0(P, \alpha, g)$ when we wish to emphasize the dependence on the contract parameters of the value to the policyholder.

The process $\{V_i\}$ represents the value of the combined annuity plus GMWB rider contract at each time point just after the deduction of fees and withdrawals. By the Markov property we have $V_i = v(i, W_i)$, where $v : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is

$$v(i, x) = \begin{cases} x & i = N, \\ [G + pv(i+1, w(xu)) + qv(i+1, w(xd))]e^{-\bar{r}} & i < N, \end{cases} \quad (7)$$

and $w : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$w(x) = \max\{xe^{-\bar{\alpha}} - G, 0\}. \quad (8)$$

Note that $\{e^{-\bar{r}i}V_i + Ga_{\bar{i}}\}_{0 \leq i \leq N}$ is a (\mathbb{Q}, \mathbb{F}) martingale for all α .

As in Hyndman and Wenger [15] we define the fair fee rate as follows.

Definition 3. A fair fee rate is a rate $\alpha^* \geq 0$ such that

$$V_0(P, \alpha^*, g) = P. \quad (9)$$

There is no closed form solution for α^* . However, as in Hyndman and Wenger [15], we are able to prove the existence and uniqueness of the fair fee rate by showing that the value V_0 is continuous and monotone as a function of α . However, in a finite probability space $\mathbb{Q}(W_N > 0) = 0$ for sufficiently large α . Consequently strict monotonicity holds only on a bounded interval.

Lemma 4. For all fixed $(i, x) \in \mathcal{I}_{N-1} \times \mathbb{R}_{++}$, the contract value function $v(i, x)$, defined by (7), as a function of α is continuous for $\alpha \geq 0$ and strictly decreasing on $[0, b^{x,i})$ where

$$b^{x,i} := \min\{\alpha \geq 0 : W_N^{x,i} = 0 \text{ a.s.}\} < \infty.$$

Further, if (i, x) satisfies

$$x > G \sum_{j=1}^{N-i} d^j \quad (10)$$

then $b^{x,i} > 0$, otherwise $b^{x,i} = 0$. For $\alpha \geq b^{x,i}$, $v(x, i) = Ga_{\overline{N-i}}$.

Proof. See Appendix A. □

In particular, equation (10) holds for $(i, x) = (0, P)$ since $G = P/N$ and $d < 1$. The existence and uniqueness of α^* is discussed in the next theorem.

Theorem 5. Under Assumption 2 there exists a unique $\alpha^* \in [0, b^{P,0})$ such that $V_0(P, \alpha^*, g) = P$.

Proof. See Appendix A. □

Remark 2. For $r = 0$ we have $V_0(P, \alpha, g) = P$ for all $\alpha \geq b^{P,0}$. Thus $r > 0$ is a necessary condition to ensure uniqueness of α^* .

From Lemma 4 we may iteratively solve for the fair fee using the bisection method provided we have a method for calculating the value V_0 as a function of α . We shall discuss the technical details of this process and computational challenges after consideration of the insurer's valuation perspective, hedging, and the extension of the model to include lapses.

2.2.2 Insurer valuation perspective

The insurer may consider the guarantees embedded in the variable annuity contract as separate products. From this point of view it is necessary to consider the time at which the account value hits zero and subsequent payments to the policyholder are drawn from the guarantee. Define the discrete-time analogue of the trigger time of the continuous model considered by Milevsky and Salisbury [19] as follows.

Definition 6. In the binomial model, the trigger time τ is defined as the stopping time

$$\tau(\omega_1 \dots \omega_N) := \inf\{i \geq 1; W_i(\omega_1 \dots \omega_i) = 0\},$$

where $\inf(\emptyset) = \infty$. For any fixed sequence $\bar{\omega}_i$ and for any $k \leq i$ we write $\tau(\bar{\omega}_i) \leq k$ if $(\bar{\omega}_i \omega_{i+1} \dots \omega_N) \in \{\tau \leq k\}$ for all possible paths $(\bar{\omega}_i \omega_{i+1} \dots \omega_N)$, where $\omega_j \in \{u, d\}$ for all $i+1 \leq j \leq N$.

It is convenient to define the respective non-decreasing sequences of stopping times $\{\tau_i\}_{i=0,1,\dots,N}$ and $\{\bar{\tau}_i\}_{i=0,1,\dots,N}$ with $\tau_i := \tau \vee i$ and $\bar{\tau}_i := \tau_i \wedge N$ for $i \in \mathcal{I}_N$. For $0 \leq j \leq i \leq N$ and $k \in \{i, i+1, \dots, N\} \cup \{\infty\}$, by the Markov property of $\{W_i\}$ we have

$$\mathbb{Q}(\tau_j = k | \mathcal{F}_i) = H(i, j, k, W_{i-}), \quad (11)$$

where

$$H(k \wedge N, j, k, x) = \begin{cases} \mathbf{1}_{\{x > 0, w(x)=0\}} & k \leq N, \\ \mathbf{1}_{\{w(x) > 0\}} & k = \infty, \end{cases}$$

and for $i \vee 1 \leq l < k \wedge N$

$$H(l, j, k, x) = \begin{cases} pH(l+1, j, k, w(x)u) + qH(l+1, j, k, w(x)d) & x > 0, \\ 0 & x = 0. \end{cases}$$

For $i = 0$, we have

$$H(0, 0, k, x) = pH(1, 0, k, xu) + qH(1, 0, k, xd).$$

If $\tau = \infty$ the contract matures with a positive account value at time $N\delta t = T$ and the option is not exercised, that is the guarantee expires worthless.

Since the value processes at each time point are ex-fees and ex-withdrawals, the component $(G - W_{\tau-}e^{-\bar{\alpha}}) \geq 0$ is the rider payment made immediately at trigger time. For any period i , the net rider payout at time $i\delta t$ is

$$(G - W_{i-}e^{-\bar{\alpha}})^+ - W_{i-}(1 - e^{-\bar{\alpha}}). \quad (12)$$

Therefore, by the risk neutral pricing formula the value at time i of the rider value process is given by

$$\begin{aligned} U_i &= E_{\mathbb{Q}} \left[\sum_{j=i+1}^N e^{-\bar{r}(j-i)} \left[(G - W_{j-}e^{-\bar{\alpha}})^+ - W_{j-}(1 - e^{-\bar{\alpha}}) \right] | \mathcal{F}_i \right] \\ &= E_{\mathbb{Q}} \left[\left(G - W_{\bar{\tau}_i-}e^{-\bar{\alpha}} \right) e^{-\bar{r}(\bar{\tau}_i-i)} \mathbf{1}_{\{i+1 \leq \bar{\tau}_i\}} + \sum_{m=\bar{\tau}_i+1}^N Ge^{-\bar{r}(m-i)} - \sum_{m=i+1}^{\bar{\tau}_i} e^{-\bar{r}(m-i)} W_{m-}(1 - e^{-\bar{\alpha}}) | \mathcal{F}_i \right] \end{aligned} \quad (13)$$

for $i \in \mathcal{I}_{N-1}$. The terminal value is $U_N = 0$. Note that equation (13) is the discrete-time binomial model analogue of equation (10) in Hyndman and Wenger [15].

By the Markov property for $\{W_i\}$ we have $U_i = u(i, W_i)$, where $u : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined by¹

$$u(i, x) = \begin{cases} 0 & i = N, \\ e^{-\bar{r}}[pu^-(i+1, xu) + qu^-(i+1, xd)] & 0 \leq i < N, \end{cases} \quad (14)$$

where $u^- : \mathcal{I}_N^+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined by

$$u^-(i, x) = u(i, w(x)) + (G - xe^{-\bar{\alpha}})^+ - x(1 - e^{-\bar{\alpha}}), \quad (15)$$

and $w(x)$ is provided by (8). The function $u^-(i, x)$ represents the rider value at time point i cum-fees and cum-withdrawals, where x is the AV before fees and withdrawals are deducted.

Since the policyholder and insurer valuation equations (5) and (13) are the respective discrete-time versions of equations (7) and (10) of Hyndman and Wenger [15] we expect that the relationship between the policyholder and insurer valuation perspectives carries over from the continuous time case. That is, we expect in the binomial model that the value of the complete contract can also be decomposed as the sum of the value of the account value and the value of the guarantee. Indeed, this can be shown directly from (5) and (13). We provide an alternative proof applying backward induction to the functions $v(i, x)$ and $u(i, x)$.

Theorem 7. *Under Assumption 2, for all $\alpha \geq 0$ we have*

$$V_i = U_i + W_i$$

for all $i = 0, 1, \dots, N$.

Proof. See Appendix A. □

One advantage of the discrete-time binomial framework is that it allows us to give explicit hedging strategies for replicating contingent claims. We next discuss hedging the GMWB rider.

2.3 Hedging

Consider the no-hedging strategy where fee revenues are invested at rate r in a money market account and at time τ , if $\tau < \infty$, the rider payoff is paid from this account. The $\mathcal{F}_{\bar{\tau}_0}$ -measurable random variable $\mathcal{C}_{\bar{\tau}_0}$ measures the total cost of the rider to the insurer over the contract lifespan, discounted to time zero, when hedging is not used. Denote the periodic fees received at time-step i by F_i . Then we have, by the definition of the contract, that $F_i := W_{i-}(1 - e^{-\bar{\alpha}})$ for $i \in \mathcal{I}_N^+$ and $F_0 = 0$. We have

$$\mathcal{C}_{\bar{\tau}_0} = e^{-\bar{r}\bar{\tau}_0} \left[\left(G - (W_{\bar{\tau}_0^-})e^{-\bar{\alpha}} \right)^+ + Ga_{N-\bar{\tau}_0} - \sum_{i=1}^{\bar{\tau}_0} F_i \times e^{\bar{r}(\bar{\tau}_0-i)} \right].$$

Note that $U_0 = E_{\mathbb{Q}}[\mathcal{C}_{\bar{\tau}_0}]$, but we are concerned with the pathwise results of $\mathcal{C}_{\bar{\tau}_0}$ in relation to the outcomes resulting from a dynamic hedging strategy.

The insurer establishes a hedging portfolio, which attempts to replicate the rider so that any rider claims can be fully paid out by the portfolio. The party managing the rider risk does not have access to the account value funds to mitigate any risk, rather the only sources of revenue are the rider fees. Denoting the replicating portfolio by $\{X_i\}$, the objective is to have $X_i = U_i$ for all i in a pathwise manner.

Define the adapted portfolio process $\{\Delta_i\}_{0 \leq i \leq N-1}$. On each time interval $[i\delta t, (i+1)\delta t)$ until maturity and for all outcomes the replicating portfolio maintains a position of $\Delta_i(\omega_1 \dots \omega_i)$ units in S . Using the Markov property of $\{W_i\}$ we define $\Delta_i := \Delta(i, W_i, S_i)$, where $\Delta : \mathcal{I}_{N-1} \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is given by

¹There is some abuse of notation with u referring both to the up-movement in the binomial model and to the rider value function. However, it is always clear from the context whether we are referring to the constant value or to the rider value function.

$$\Delta(i, x, y) = \frac{u^-(i+1, ux) - u^-(i+1, dx)}{uy - dy}. \quad (16)$$

This indicates that $\Delta_i = 0$ for $\tau \leq i \leq N-1$ as no uncertainty remains. By the nature of the rider as an embedded put-like option, Δ will always take non-positive values corresponding to short positions in S . Any positive (negative) portfolio cash balance is invested in (borrowed from) the money market at rate r .

Beginning with initial capital $X_0 = x_0 \in \mathbb{R}$, the replicating portfolio $\{X_i\}$ follows

$$X_i = (X_{i-1} - \Delta_{i-1}S_{i-1})e^{\bar{r}} + \Delta_{i-1}S_i + F_i - (G - W_{i-}e^{-\bar{\alpha}})^+ \quad (17)$$

for $i \in \mathcal{I}_N^+$. Over any period the change in the portfolio value of $(X_i - X_{i-1})$ consists of the sum of four components: a) the return in the money market earned on both the prior portfolio balance and the proceeds from the shorted stock $(X_{i-1} - \Delta_{i-1}S_{i-1})(\exp(\bar{r}) - 1)$; b) the capital gain or loss on the shorted stock $(S_i - S_{i-1})\Delta_{i-1}$; c) the end of period rider fees F_i ; and d) the negative of that period's rider claim (if any), paid at the end of the period and given by $(G - W_{i-} \exp(-\bar{\alpha}))^+$. Note that if the static hedging strategy $\Delta \equiv 0$ is used then $X_N \exp(-\bar{r}N) = -C_{\tau_0}$. That is, we obtain the no-hedging result.

Similar to Shreve [23, Theorem 2.4.8] the portfolio process given by (16) replicates the rider value. The proof is omitted as we shall prove a more general result after we generalize the model to include lapses.

Theorem 8. *Under Assumption 2, if the fee α is charged and the initial capital is $x_0 = U_0(P, \alpha, g)$, then an insurer who maintains the replicating portfolio X_i by following the portfolio process prescribed by (16) will be fully hedged. That is,*

$$X_i = U_i$$

for $i \in \mathcal{I}_N$.

Remark 3. *In particular, if $\tau \leq N$ then $X_\tau = G \times a_{\overline{N-\tau}|}$. When α^* is charged we have $U_0 = 0$ and no initial capital is required for the replicating portfolio. The rider is different from standard financial options in that there is no upfront cost to finance the hedge but rather it is self-financed through periodic contingent fees. If the fee charged is not the fair fee ($\alpha \neq \alpha^*$), then the insurer must make an initial deposit to the hedging portfolio if $\alpha < \alpha^*$ or may consume from the portfolio at time zero if $\alpha > \alpha^*$. The insurer can justify a lower fee by either depositing capital into the portfolio and selling the policy at a loss or by charging an initial fee per unit premium at time zero to the insured.*

3 Optimal Stopping and Surrenders

We next extend the binomial pricing model to include the possibility of early surrenders by modifying Assumption 2 to include the following assumption.

Assumption 9. *Under the static withdrawal strategy the policyholder receives $G = gP\delta t$ each time period. We set $T := 1/g$. At the end of each period the pro-rated rider fee is first deducted and then the periodic withdrawal is subtracted. We restrict $r > 0$ and denote $\bar{r} := r\delta t$ and $\bar{\alpha} := \alpha\delta t$. Surrenders occur at the end of any time period, after the fees and withdrawals have been deducted. For valuation purposes, the end of period time point is considered ex-post fees and withdrawals but ex-ante surrenders.*

Let $k^a : \{0, 1, \dots, T\} \mapsto [0, 1]$ be the non-increasing function describing the surrender charge schedule, satisfying $k_0^a > 0$ and $k_T^a = 0$. The surrender charge rate k_i^a is applied for surrenders during time $[i, i + 1)$. We denote the corresponding function for the surrender charge rate upon surrender at the end of period i by $k : \{0, 1, \dots, N\} \rightarrow [0, 1]$. Then $k_i = k_{\lfloor i\delta t \rfloor}^a$. Similar to the continuous-time model of Hyndman and Wenger [15] for all $i \in \mathcal{I}_N$ we have

$$V_i = \max_{\eta \in \mathbb{L}_i} V_i^\eta = \max_{\eta \in \mathbb{L}_i, \bar{\tau}_i} V_i^\eta, \quad (18)$$

where

$$V_i^\eta = E_Q \left[G a_{\overline{\eta-i}} + W_\eta (1 - k_\eta) e^{-\bar{r}(\eta-i)} | \mathcal{F}_i \right], \quad (19)$$

\mathbb{L}_i is the set of \mathbb{F} -adapted stopping times taking values in $\{i, i + 1, \dots, N\}$, and $\mathbb{L}_{i, \bar{\tau}_i}$ is the set of \mathbb{F} -adapted stopping times taking values in $\{i, i + 1, \dots, N\}$ subject to the constraint $\eta < \bar{\tau}_i$ or $\eta = N$. Recall that $\bar{\tau}_i$ is the trigger time assuming no lapses.

With the objective of classifying the optimal surrender policy we introduce some notation. For any $0 \leq i \leq N$, define a rescaled filtration $\mathbb{F}^i = \{\mathcal{F}_j^i := \mathcal{F}_{j+i}; 0 \leq j \leq N - i\}$. For any $\eta \in \mathbb{L}_i$ define

$$Y^{\eta, i} := \left\{ Y_j^{\eta, i} = e^{-\bar{r}((j+i) \wedge \eta)} V_{(j+i) \wedge \eta}^\eta + G a_{\overline{(j+i) \wedge \eta}} \right\}_{0 \leq j \leq N-i}, \quad (20)$$

then $Y^{\eta, i}$ is a $(\mathbb{Q}, \mathbb{F}^i)$ martingale. Define the surrender policy $\tilde{\eta}$ by

$$\tilde{\eta}_i := \min\{j \geq i; V_j = W_j(1 - k_j)\} \leq N. \quad (21)$$

A policyholder following the surrender strategy given by equation (21) lapses at the first time valuation of the contract, from the policyholder's perspective, is equal to the account value less the surrender charge. This is similar to the classical result from American contingent claims theory which gives that $\tilde{\eta}_i$ is optimal in the sense that $V_i = V_i^{\tilde{\eta}_i}$ (proving this in our context is straightforward based on Duffie [10, p.35] but requires (20)). That is, $\tilde{\eta}_i$ is an optimal surrender policy for the insured to follow going forth from time $i\delta t$, given the current market state and no prior surrender.

The backward induction (risk-neutral pricing) algorithm is constructed to evaluate V on a binomial tree. By the Markov property for $\{W_i\}$ we have $V_i = v(i, W_i)$, where $v : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given recursively as

$$\begin{cases} v(N, x) = x(1 - k_N) = x, \\ v(i, x) = \max\{(G + pv(i + 1, w(ux)) + qv(i + 1, w(dx)))e^{-\bar{r}}, x(1 - k_i)\}. \end{cases}$$

When solving for α^* we may write

$$v(0, P) = [G + pv(1, w(uP)) + qv(1, w(dP))]e^{-\bar{r}}$$

since $k_0 > 0$.

Consider the rider value U by extending equation (13) to incorporate the option to surrender and receive the payoff $k_\eta W_\eta$ at surrender. Then, at time i the rider value is given by

$$U_i := \max_{\eta \in \mathbb{L}_i, \bar{\tau}_i} U_i^\eta \quad (22)$$

where

$$U_i^\eta = E_Q \left[\sum_{j=i+1}^{\eta} e^{-\bar{r}(j-i)} [(G - W_{j-} e^{-\bar{\alpha}})^+ - W_{j-} (1 - e^{-\bar{\alpha}})] - e^{-\bar{r}(\eta-i)} k_\eta W_\eta | \mathcal{F}_i \right]$$

using the convention that $\sum_{j=i+1}^i (\cdot) = 0$. Note that equation (22) is the discrete-time analogue of the rider value in the continuous-time model given in equation (14) of Hyndman and Wenger [15].

The value of the option to surrender, L , is the difference between the rider value when lapses are allowed and the rider value without lapses. That is, define $L_i := U_i - U_i^{NL} \geq 0$, where U_i^{NL} is the rider value in the no-lapse case (13). Then at time i the value of the option to lapse is given by

$$L_i = \max_{\eta \in \mathcal{L}_i, \bar{\tau}_i} L_i^\eta, \quad (23)$$

where

$$L_i^\eta = E_{\mathbb{Q}} \left[\sum_{j=\eta+1}^N e^{-\bar{r}(j-i)} [W_{j-} (1 - e^{-\bar{\alpha}}) - (G - W_{j-} e^{-\bar{\alpha}})^+] - e^{-\bar{r}(\eta-i)} k_\eta W_\eta | \mathcal{F}_i \right].$$

Note that equation (23) is the discrete time analogue of the value of the option to lapse in the continuous-time model given by equation (15) of Hyndman and Wenger [15].

Write $U_i = u(i, W_i)$, where $u : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}$ is recursively defined by

$$\begin{cases} u(N, x) = -k_N x = 0, \\ u(i, x) = \max\{e^{-\bar{r}}[pu^-(i+1, ux) + qu^-(i+1, dx)], -k_i x\}, \end{cases}$$

and $u^- : \mathcal{I}_N^+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ follows

$$u^-(i, x) = u(i, w(x)) + (G - xe^{-\bar{\alpha}})^+ - x(1 - e^{-\bar{\alpha}}). \quad (24)$$

Denoting the rider value function in the no-lapse model from (14) by $u^{NL}(i, x)$, we have $L_i = l(i, W_i)$, where $l : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$\begin{cases} l(N, x) = -k_N x = 0, \\ l(i, x) = \max\{e^{-\bar{r}}(pl(i+1, w(ux)) + ql(i+1, w(dx))), -u^{NL}(i, x) - k_i x\}. \end{cases}$$

Note that $u^{NL}(i, 0) \geq 0$ which implies the boundary condition $l(i, 0) = 0$. Once the rider is triggered, early surrender is suboptimal since any remaining guarantee is forfeited upon surrender.

In the case of lapses we may extend Theorem 7 to decompose the value of the complete contract into the sum of the account value and the value of the guarantee.

Theorem 10. *Under Assumption 9, for all $\alpha \geq 0$ and for all $i \in \mathcal{I}_N$, we have*

$$V_i = U_i + W_i, \quad (25)$$

or equivalently

$$V_i = L_i + U_i^{NL} + W_i. \quad (26)$$

Proof. Equation (25) can be proved using backward induction on the recursive functions v and u , similar to Theorem 7. We omit the details. \square

Note that Theorem 10 is the discrete-time analogue of the continuous time decomposition given in Hyndman and Wenger [15, Theorem 7].

As in the no-lapse case an advantage of the discrete-time binomial model is that we are easily able to hedge the guarantee.

3.1 Hedging with lapses

We next extend the standard hedging results for American derivatives (see Shreve [23, Theorem 4.4.4]) by incorporating the complication of the periodic revenues and rider claims. We show that the insurer can perfectly hedge the rider risk by maintaining the appropriate replicating portfolio. The adapted portfolio process $(\Delta_i)_{0 \leq i < N}$ remains unchanged from (16), except that $u^-(i, x)$ is given by (24). Furthermore, the insurer may have positive consumption under suboptimal surrender behaviour.

Define the consumption process $C = \{C_i\}_{0 \leq i < N}$ by $C_i := c(i, W_i)$ where $c : \mathcal{I}_{N-1} \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$c(i, x) := v(i, x) - [pv(i+1, w(ux)) + qv(i+1, w(dx)) + G]e^{-\bar{r}} \geq 0. \quad (27)$$

The consumption process $\{C_i\}$ represents the additional cash flow received each time a policyholder behaves sub-optimally by not surrendering. We can explicitly classify suboptimal behaviour by defining a sequence of stopping times.

With $\tilde{\eta}_i$ defined as in equation (21) let $\tilde{\eta}^0 := \tilde{\eta}_0$ and for $1 \leq j \leq m$ we denote

$$\tilde{\eta}^j = \tilde{\eta}_{z_j},$$

where $z_0 = 0$, $z_j = (\tilde{\eta}^{j-1} + 1) \wedge N$, and $m = \min\{i; \tilde{\eta}^i = N \text{ a.s.}\}$. Then we may characterize, in terms of $\{\tilde{\eta}^j\}$, precisely when C will be strictly positive. We have $C_{\tilde{\eta}^j} > 0$ for all $0 \leq j < M := \min\{b; z_b = N\} \leq m$, where M is a random variable. Otherwise $C_i = 0$.

There is a fine distinction between $C_{\tilde{\eta}^j}$ and $L_{\tilde{\eta}^j}$ for all $j < M$. Consider the two surrender strategies of $\tilde{\eta}^{j+1}$ and $\eta = N$. The first strategy corresponds to surrendering at the next best time after $\tilde{\eta}^j$ and the latter strategy is equivalent to never surrendering early. Then $C_{\tilde{\eta}^j} = V_{\tilde{\eta}^j} - V_{\tilde{\eta}^j}^{\tilde{\eta}^{j+1}}$ but $L_{\tilde{\eta}^j} = V_{\tilde{\eta}^j} - V_{\tilde{\eta}^j}^{NL}$. At any time when it is optimal to surrender immediately, C provides the marginal value from surrendering now instead of at the next optimal time, whereas L is the marginal value from acting now instead of at maturity.

By Theorem 7 and Theorem 10 it follows that $V_{\tilde{\eta}^j}^{\tilde{\eta}^{j+1}} = U_{\tilde{\eta}^j}^{\tilde{\eta}^{j+1}} + W_{\tilde{\eta}^j}$ and $V_{\tilde{\eta}^j} = U_{\tilde{\eta}^j} + W_{\tilde{\eta}^j}$. Therefore C can be written in terms of U as

$$c(i, x) = u(i, x) - [pu^-(i+1, ux) + qu^-(i+1, dx)]e^{-\bar{r}}. \quad (28)$$

Beginning with $X_0 = x_0$, the replicating portfolio is constructed recursively forward in time taking into consideration fee revenues, consumption, and rider claim payments. For all $i \in \mathcal{I}_N^+$ we have

$$X_i = [X_{i-1} - \Delta_{i-1}S_{i-1} - C_{i-1}]e^{\bar{r}} + \Delta_{i-1}S_i + F_i - (G - W_{i-1}e^{-\bar{\alpha}})^+. \quad (29)$$

Theorem 11. *Under Assumption 9, if the initial capital is $x_0 = U_0$, then an insurer who maintains the replicating portfolio X_i defined by (29) and liquidates the portfolio either upon early surrender (if any) or at time point N will be fully hedged throughout the contract lifespan. That is, for all $i \in \mathcal{I}_N$ and all surrender strategies*

$$X_i = U_i.$$

Proof. See Appendix A. □

Remark 4. *Assuming the insured follows the optimal surrender strategy $\tilde{\eta}_0$, then $X_{\tilde{\eta}_0} = U_{\tilde{\eta}_0}$ and on $\{\tilde{\eta}_0 < \bar{\tau}_0\}$ we have that $X_{\tilde{\eta}_0} = U_{\tilde{\eta}_0} = -k_{\tilde{\eta}_0}W_{\tilde{\eta}_0}$, whereas $X_N = U_N = 0$ on $\{\tilde{\eta}_0 = N\}$. Under this strategy there is no consumption. However, if the insured allows the first optimal surrender time $\{\tilde{\eta}_0 < \bar{\tau}_0\}$ to elapse, then the insurer will consume $C_{\tilde{\eta}_0}$ and the remaining portfolio is still sufficient to hedge the contract over the remaining lifespan. If the insured allows the next optimal surrender time $\{\tilde{\eta}_0 < \tilde{\eta}^1 < \bar{\tau}_0\}$ to elapse, if it exists, then the insurer consumes an additional $C_{\tilde{\eta}^1}$ and this continues until the earlier of trigger or time point N .*

Finally suppose the insured surrenders at a suboptimal time. For a given path $\bar{\omega}_N$, surrender occurs at a time point $i \neq \tilde{\eta}^j$ for all $0 \leq j \leq M(\bar{\omega}_N)$. Then the insured receives $W_i(1 - k_i)$ and in turn foregoes $V_i - W_i(1 - k_i) > 0$ of value. The insurer's portfolio value is $X_i + k_iW_i > 0$ and the insurer has a positive consumption. Indeed by (25) we have $V_i - W_i(1 - k_i) = U_i + W_i k_i > 0$, but $X_i = U_i$.

With the explicit recursive formulae for pricing and hedging the contract we may consider the implementation of the binomial model and its performance relative to the theoretical results presented and other modeling approaches which have appeared in the literature. We first briefly address computational considerations of the binomial model.

4 Computational Considerations

Computational applications of the binomial model for the GMWB rider face two specific challenges. The binomial tree for the account value process is non-recombining and the riders have significantly longer durations in contrast to the usual European and American equity options which typically have durations not exceeding one year. The withdrawal rate g can be expected to range from 5% to 10% corresponding to maturities of 10 to 20 years. If the value processes in the binomial world is to provide an accurate approximation of the value processes in the continuous-time model of Hyndman and Wenger [15] δt must be significantly smaller than one.

The backward induction (tree) algorithm (referred to as Method A) for calculating V_0 involves arrays of size 2^N to record V_N for all nodes in the final period. In contrast, for recombining trees the array size needed is only $N + 1$. For $g = 5\%$ the binomial tree will contain $2^{20} > 10^6$ nodes in the final period with just one time-step per year. Method A requires too much memory for small values of δt .

We will show that in the no-lapse model we can directly calculate $v(i, x)$ without using trees and avoid the strain on memory capacity from storing the large arrays of data. This direct approach (Method B) uses an algorithm which loops through each path requiring minimal memory. We will see shortly that despite being able to eliminate a subset of the paths from the looping process this method is significantly slower than Method A. Although Method B enables using marginally smaller δt values, we quickly run into time constraints as the number of paths grows at $O(2^N)$.

We will then introduce an approximation method which uses the backward induction (tree) approach while easing the memory strain. This retains the flexibility to model the GMWB both with and without lapses. Further it avoids the time constraints with Method B.

The terminal AV can be expressed directly as:

$$\begin{aligned} W_N &= \max \left[\xi_N e^{-\bar{\alpha}} (\xi_{N-1} e^{-\bar{\alpha}} (\dots (\xi_2 e^{-\bar{\alpha}} (P \xi_1 e^{-\bar{\alpha}} - G) - G) \dots) - G) - G, 0 \right] \\ &= \max \left[0, P e^{-\bar{\alpha} N} \prod_{i=1}^N \xi_i - G \sum_{i=0}^{N-1} e^{-\bar{\alpha} i} \prod_{j=N-(i-1)}^N \xi_j \right], \end{aligned} \quad (30)$$

where the convention $\prod_{N+1}^N(\cdot) = 1$ is used. Applying the reversal technique from Liu [18], which is justified by the exchangeability property of the sequence $\{\xi_i\}_{i=1}^N$, and considering the reversed sequence which is equal in distribution, it follows that

$$W_N^{x,M} \stackrel{d}{=} \max \left[0, x Z_{N-M} - G \sum_{i=0}^{N-M-1} Z_i \right],$$

where $M < N$ and $\{Z_i\}$ is the account value process when there are no withdrawals, beginning with $Z_0 = 1$. In particular, with $M = 0$, $x = P$, and $G = P/N$ we obtain that V_0 can be expressed as a floating-strike Asian call option on $\{Z_i\}$ plus a term certain component, as pointed out by Liu [18].

Many of the terminal nodes in the tree for $\{W_i\}$ will be zero as a result of the periodic withdrawals, fees, and possible negative returns on S . Consider the recombining tree for $\{Z_i\}$ with $N + 1$ nodes for period N . At each node, for each path leading to it the average must be computed to calculate W_N . Suppose that for some $i \leq N$ we have $W_N = 0$ on all paths with i jumps of u and $N - i$ jumps of d . Then $W_N = 0$ for all paths with less than i jumps of u . Consequently, once we reach a node on the tree for Z such that $W_N = 0$ for all paths, no further paths need be considered.

There is an efficient permutation function in C++, *next_permutation*, which quickly loops through all distinct paths having i jumps of u and $N - i$ jumps of d . By looping through each node and its respective paths we can avoid the exponential growth in memory storage, although we show in our numerical results that the run-time will increase significantly. By (5), with $\zeta := N - m$ we can write

$$v(m, x) = Ga_{\zeta} + e^{-\bar{r}\zeta} \sum_{k=0}^{A_0} p^{\zeta-k} q^k \sum_{\Xi_{\zeta,k}} \left(x e^{-\bar{\alpha}\zeta} u^{\zeta-k} d^k - G \sum_{i=0}^{\zeta-1} e^{-\bar{\alpha}i} \prod_{j=1}^i \omega_j \right)^+, \quad (31)$$

where $\Xi_{\zeta,k}$ is the set of $\binom{\zeta}{k}$ unique permutations of a path with $\zeta - k$ up and k down movements and A_0 is the first value of k for which the summand produces zero.

Hull and White [14] developed an approximation method to value path-dependent financial options on a binomial lattice in a more efficient manner. The key idea is to use only a representative set of averages at each node and apply linear interpolation in the backwards induction scheme. Costabile et al. [7] discuss several drawbacks of the Hull and White [14] method and propose a different approximation method and in particular provide the details for pricing fixed-strike European and American Asian call options. Numerical results show convergence for European Asian calls while American Asian calls do not perform as well, converging at a much slower rate. The method is easily modified for any option payoff which depends on a valid function of the asset price path.

The options considered by Costabile et al. [7] have significantly shorter maturities compared to the GMWB riders. The method reduces the number of contract values considered in the backwards induction scheme from $O(2^N)$ to $O(N^4)$. In our work, memory constraints limited the number of time steps in the binomial trees to $N = 28$ but with this method we can consider up to $N = 128$ time-steps. We briefly describe the approximation method applied to GMWBs with lapses but refer the reader to Costabile et al. [7] for more details on the scheme.

Using equation (30) we can rewrite the value of the contract to the policyholder given by equation (18) as

$$V_0 = \max_{\eta \in \mathbb{L}_0} E_Q \left[Ga_{\bar{\eta}} + P \max \left(Z_{\eta} \left(1 - \frac{1}{N} \sum_{i=1}^{\eta} \frac{1}{Z_i} \right), 0 \right) (1 - k_{\eta}) e^{-\bar{r}\eta} \right], \quad (32)$$

where

$$Z_n = \prod_{i=1}^n e^{-\bar{\alpha}\xi_i} = e^{-\bar{\alpha}n} \frac{S_n}{S_0}.$$

Therefore,

$$V_i = v(i, Z_i, \sum_{j=1}^i Z_j^{-1}),$$

where $v : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined recursively backward in time by

$$v(N, x, y) = P \max \left(x \left(1 - \frac{1}{N} y \right), 0 \right)$$

for $i = N$ and

$$\begin{aligned} v(i, x, y) = \max & \left[\left[G + pv(i+1, xue^{-\bar{\alpha}}, y + (xue^{-\bar{\alpha}})^{-1}) \right. \right. \\ & \left. \left. + qv(i+1, xde^{-\bar{\alpha}}, y + (xde^{-\bar{\alpha}})^{-1}) \right] e^{-\bar{r}}, x \left(1 - \frac{1}{N} y \right) (1 - k_i) \right] \end{aligned}$$

for $0 \leq i < N$.

Let (i, j) denote the node reached by j up-movements and $(i - j)$ down-movements in the recombining tree for Z . We write $z(i, j)$ for the value of Z at node (i, j) . For each node, we construct a set of $j(i - j) + 1$ representative averages, where the terminology of *average* is used even though we do not divide by $i + 1$. This set is a subset of the complete set of $\binom{i}{j}$ averages for the paths at that node. Denote the first (and lowest) element by $A(i, j, 1)$ where

$$A(i, j, 1) = \sum_{h=0}^j (ue^{-\bar{\alpha}})^{-h} + (ue^{-\bar{\alpha}})^{-j} \sum_{h=1}^{i-j} (de^{-\bar{\alpha}})^{-h}.$$

This average is taken along the path beginning with j up-movements of u and followed by $(i - j)$ down-movements of d . Excluding the initial point and terminal point we find the highest point of $\{S_i\}$ along the path (if there are more than one such points, select the first one) and substitute that node with the node directly below it in the $\{Z_i\}$ tree to obtain a new path and take its average. This is repeated $j(i - j)$ times to obtain the set $A(i, j) = \{A(i, j, k); 1 \leq k \leq j(i - j) + 1\}$. The final path considered will be the one with $(i - j)$ down-movements followed by j up-movements. None of the previous paths are allowed to be below this path.

When working with the function v on the tree for Z and applying backward induction, linear interpolation is used whenever the computed average is not in the representative set for that node. This is done by considering the two nearest elements of the set, one on each side of the computed average (see Hull and White [14] and Costabile et al. [7] for details). The scheme from Costabile et al. [7] has the benefit that linear interpolation is not needed for many of the computations of v .

For the framework in Costabile et al. [7], whether the algorithm begins with the path giving the highest average, selects paths in the described manner, and stops when the path giving the lowest average is obtained, or vice versa, the same set of averages are obtained. This symmetry is a result of the underlying asset changing by factors of u and d , where $ud = 1$. However, this symmetry does not hold in our model because the process Z changes by factors of $ue^{-\alpha}$ and $de^{-\alpha}$. For example, an up-move followed by a down-move does not return Z to its initial value. The downward trend of the Z -tree complicates the approximation algorithm. Consequently, the sets $A(i, j)$ will change depending on whether the lowest or highest path is initially considered.

5 Numerical Results: Excluding Mortality Risk

Beginning with the no-lapse case, we provide numerical results comparing our model to previous results in the literature, which excluded mortality risk, and find that even with large values for δt our simple model is a reasonable approximation of more complex models. Moreover, the discrete-time binomial model allows us to analyze the hedging results and the effect of the parameters on the losses when hedging is not implemented.

5.1 The Fair Rider Fee

The bisection algorithm is used to numerically solve for α^* given by Definition 3. Define $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ by $f(\alpha) = V_0(P, \alpha, g) - P$. Then $f(\alpha^*) = 0$ by Definition 3. We use $P = 100$ and stop iterations when $|f(\alpha)| < \epsilon^*$ where $\epsilon^* \leq 0.001$ in all our results achieving accuracy of 1×10^{-5} for a unit premium.

In the continuous-time model Milevsky and Salisbury [19] use numerical PDE techniques to solve for V_0 , corresponding to Hyndman and Wenger [15, equation. (7)], and present the fair fees for various (g, σ) combinations. In Liu [18], a discrete-time model is developed and the contract values are estimated using Monte Carlo simulation with a geometric mean strike Asian call option as a control variate. Both papers assume S is log-normally distributed. In theory we expect convergence of results for both models and our binomial model. However Liu [18] obtains results significantly lower than those of Milevsky and Salisbury [19], and concludes that the results of Milevsky and Salisbury [19] are on average 28% too high.

Table 1 provides a comparison between the results of Milevsky and Salisbury [19], Liu [18], and the binomial model. In the discrete models $\delta t = 1/\text{time-steps}$. The parameters are: $P = 100$, $g = 10\%$, $r = 5\%$, $\sigma = 20\%$, $T = 1/g = 10$. For $\delta t = 1$, results from the binomial model and Liu [18] are sufficiently close. We reach three time-steps per year under Method B, and observe that the binomial model supports the results of Liu [18].

For the same parameters Table 2 displays sample run-times (in seconds) to calculate V_0 for a single value of α . The differences may seem small for $n < 3$ and external factors also affect the run-times. However Method A is implemented in *Matlab* while Method B is implemented in C++ which is generally more efficient for identical code. Therefore, we find that Method B is significantly slower. Under Method B with $n = 3$ and $\alpha = 95.35\text{bps}$, we observe that $W_N = 0$ for all paths with less than 11 up-moves and, therefore, the bottom 10 nodes in the recombining tree for Z do not need to be evaluated. However, this simplification does not prevent the run-time from growing rapidly with n .

	M&S (2006)	Liu [18]			Binomial		
time-steps/year	continuous	1	12	4000	1	2	3
α^* (bps)	140	92.41	96.65	97.28	92.20	94.55	95.35

Table 1: Comparison of results for α^* : $g = 10\%$, $r = 5\%$, $\sigma = 20\%$

Time-Steps	Method A	Method B
	(Trees, Matlab)	(Loop, C++)
$n = 1$	7.7×10^{-4}	3×10^{-3}
$n = 2$	0.80	2.5
$n = 3$		3×10^3

Table 2: Computational time comparison (in seconds)

While the binomial model is a valuable theoretical framework for viewing the GMWB rider, it is the Asian approximation method which reveals the practical value of such a model. Implementing the Asian approximation method, we attain results up to $n = 10$. Monthly time-steps should be attainable with more efficient programming and superior hardware. The results in Table 3 imply convergence to the α^* computed by Liu [18].

Table 4 contains additional results for different g and σ values. The fair fee is increasing with both g and σ and is quite sensitive to σ . Sensitivity results have been discussed at length in the literature (see Chen et al. [6]). The return of premium guaranteed by the GMWB does not include time value of money and as g increases, the maturity decreases and V_0 increases in value for any fixed α because of the interest rate effect. Consequently α^* must increase. Our results consistently support Liu [18] at the expense of Milevsky and Salisbury [19].

In Figure 1, V_0 is plotted against α for different T values. The parameters are: $P = 100$, $r = 5\%$, $\sigma = 20\%$, $\delta t = 1$, and $g = (1/T)$. The fair fee is the point of intersection between the horizontal line $V_0 = 100$ and the curves. When the curves are plotted over the wider range $[0, 0.05]$ the linearity resemblance seen on $[0, 0.01]$ disappears and the curves have a more pronounced convex shape. As α increases, the likelihood of trigger rises but the decrease in the expected discounted terminal account value is less sensitive for sufficiently large α .

It is important to consider the sensitivity of V_0 to α in a neighbourhood around α^* , for a given set of parameters. Figure 1 reflects the changing sensitivity for different values of T . For the parameters in Table 1, the binomial method with $\delta t = 2$ gives $V_0(100, 140 \text{ bps}, 10\%) = 98.02$ and it can be deceptive to only look at α^* . The objective is to solve for the fair fee and in our pricing framework, charging a different fee leads to arbitrage no matter the size of $|\alpha - \alpha^*|$. However, in the presence of real world constraints such as imperfect models, market frictions, and sub-rational policyholder behaviour small pricing errors may not lead to arbitrage and it is crucial consider price sensitivity in

n	1	2	3	5	7	9	10
α^* (bps)	92.30	94.64	95.40	96.05	96.33	96.48	96.54
$V_0(\alpha = 97.3)(\$)$	99.767	99.880	99.917	99.945	99.958	99.965	99.967

Table 3: Asian approximation results

(α^*, bps)		$\sigma = 20\%$			$\sigma = 30\%$		
$g\%$	T	MS ^a	L ^b	B ^c	MS ^a	L ^b	B ^c
5	20	37	28.5	27.1(1)	90	76.5	74.8(1)
6	16.67	54	40.6	38.7(1)	123	103.7	101.5(1)
7	14.29	73	53.8	51.3(1)	158	132.3	129.4(1)
8	12.5	94	n/a	64.6(1)	194	n/a	158.3(1)
9	11.11	117	n/a	80.1(2)	232	n/a	189.3(2)
10	10	140	96.7	94.6(2)	271	221.2	219.1(2)

^a Milevsky and Salisbury [19] ^b Liu [18] with $n = 12$

^c Binomial with n in parentheses

Table 4: Comparison with previous results for α^* , ($r = 5\%$)

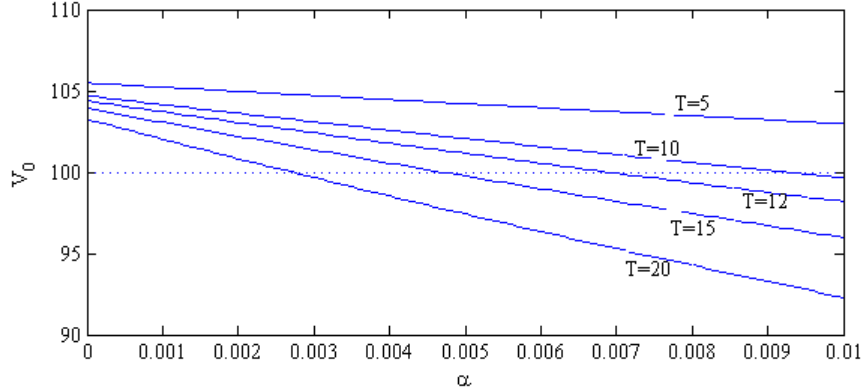


Figure 1: Plotting V_0 as a function of α for varying T . Parameters are: $r = 5\%$, $\sigma = 20\%$, and $g = 1/T$.

addition to finding α^* .

5.2 Distribution of the Trigger

Milevsky and Salisbury [19] numerically solve the Kolmogorov backward equation for $\mathbb{P}(\tau \leq T)$ and provide results for different combinations of (μ, σ) with the parameters $g = 7\%$ and $\alpha = 40\text{bps}$. To avoid fractional years, we set $T = 14$ and $g = 7.14\%$. As shown in Table 5, the binomial model with just $n = 2$ produces probabilities close to Milevsky and Salisbury [19]. The accuracy improves with increasing σ .

In Milevsky and Salisbury [19], S_t is modelled by geometric Brownian motion

$$dS_t = \mu S_t dt + \sigma S_t dB'_t,$$

where B'_t is \mathbb{P} -Brownian motion. Then with

$$r_T^s := \ln\left(\frac{S_T}{S_0}\right) = \left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma B'_T,$$

	$\sigma = 10\%$				$\sigma = 15\%$				$\sigma = 18\%$				$\sigma = 25\%$			
	M&S		Binomial		M&S		Binomial		M&S		Binomial		M&S		Binomial	
	$\delta t = 0.5$	$\delta t = 1$	$\delta t = 0.5$	$\delta t = 1$	$\delta t = 0.5$	$\delta t = 1$	$\delta t = 0.5$	$\delta t = 1$	$\delta t = 0.5$	$\delta t = 1$	$\delta t = 0.5$	$\delta t = 1$	$\delta t = 0.5$	$\delta t = 1$	$\delta t = 0.5$	$\delta t = 1$
$\mu = 4\%$	19.0%	16.0%	15.2%	31.4%	31.1%	30.9%	37.8%	38.2%	38.2%	49.9%	50.8%	50.6%				
$\mu = 6\%$	7.0%	4.5%	3.6%	18.5%	17.8%	16.9%	25.5%	25.3%	25.0%	39.6%	40.5%	40.2%				
$\mu = 8\%$	1.7%	0.7%	0.3%	9.3%	8.2%	7.4%	15.5%	15.0%	14.5%	30.5%	30.8%	30.4%				
$\mu = 10\%$	0.3%	0.0%	0.0%	4.1%	3.1%	2.3%	8.6%	7.8%	7.1%	22.2%	22.2%	21.7%				
$\mu = 12\%$	0.04%	0.0%	-	1.6%	0.9%	0.4%	4.4%	3.5%	2.7%	15.5%	15.2%	14.4%				

Table 5: $\mathbb{P}(\tau < \infty)$: comparing binomial model to continuous time model from Milevsky and Salisbury [19]

we have $\mathbb{E}_{\mathbb{P}}[r_T^s] = (\mu - \frac{1}{2}\sigma^2)T$ and $\text{Var}_{\mathbb{P}}[r_T^s] = \sigma^2 T$. In the binomial model we set

$$\begin{aligned} u &= e^{\sigma\sqrt{\delta t}}, \\ d &= e^{-\sigma\sqrt{\delta t}}, \\ \tilde{p} &= \frac{1}{2} + \frac{1}{2} \left(\mu - \frac{1}{2}\sigma^2 \right) \frac{1}{\sigma} \sqrt{\delta t}. \end{aligned}$$

Note that $\tilde{p} < 1$ holds only if $\mu < \frac{1}{2}\sigma^2 + \sigma \frac{1}{\sqrt{\delta t}}$. For $\delta t = 1$ this condition is violated for $\sigma = 10\%$ and $\mu = 12\%$.

In general, the probability mass function of τ with respect to \mathbb{P} can be calculated in the binomial model using equation (11), where

$$\mathbb{P}(\tau = i) = H(0, 0, i, P)$$

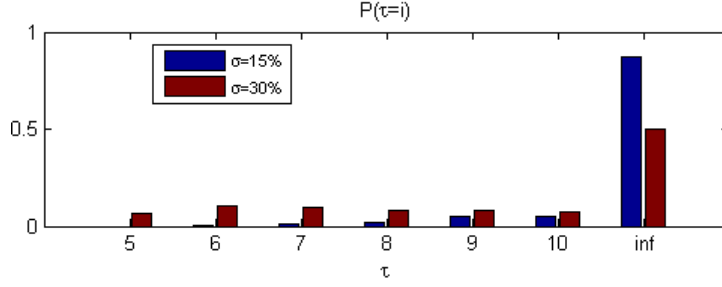
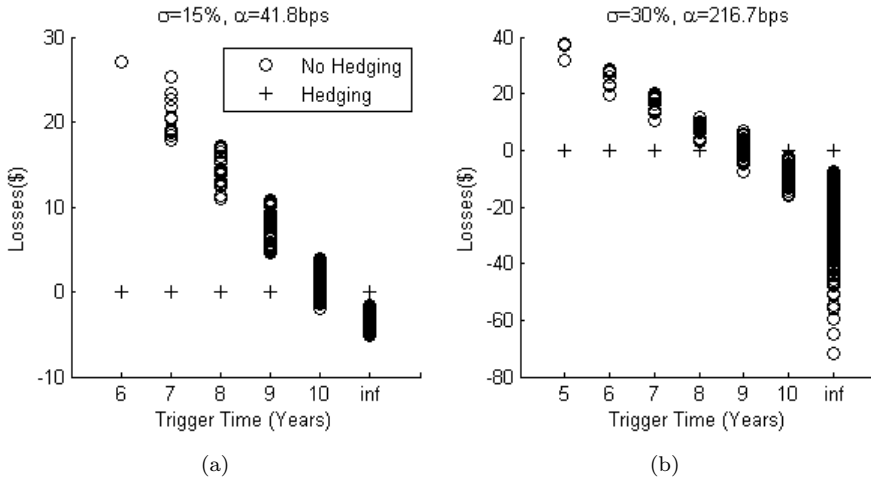
for $i \in \{1, 2, \dots, N, \infty\}$. Of course, p must be replaced with \tilde{p} .

Remark 5. Applying equation (11) to calculate the trigger probabilities with two time-steps a year, 2^{28} paths need to be evaluated and we run into capacity issues. For $\delta t = 0.5$, we use the approach of equation (31) except that rather than working with $\exp(-rT)W_T$, we use the indicator function $\mathbf{1}_{\{W_T=0\}}$ remembering to take account of the probabilities for the lower nodes with more than A_0 down movements.

5.3 Comparison of Hedging and No Hedging

We investigate the impact of volatility on the fees, triggers and losses. The parameters are: $g = 10\%$, $T = 10$, $P = 100$, and $\delta t = 1$. The risk free rate r is 5% and the drift term μ of the underlying asset is 7.5%. We consider $\sigma = 15\%$ and $\sigma = 30\%$. The respective fair fees α^* are 41.8bps and 216.7bps. The probability mass function for τ under the physical measure is displayed in Figure 2. Recall that $\tau = \infty$ when $W_T > 0$. The two σ values were selected to magnify the interaction between volatility, the trigger time distribution and consequently the rider payouts. Higher volatility implies more adverse market returns and a greater likelihood of early trigger. An additional effect on trigger comes from the rider fee. The fee rate is very sensitive to volatility and the fees drag down the account value further, resulting in more frequent early trigger times.

We consider the strategies of no hedging and dynamic delta hedging prescribed in Subsection 2.3. Define $\Pi := \exp(-\bar{r}N)X_N$ to be the discounted profit. When Δ follows the prescribed portfolio process (16) we obtain the hedging profit, Π^H . If $\Delta \equiv 0$ we obtain the profit under no hedging, Π^{NH} . The superscripts are omitted when it is clear which profits we are analyzing. Figure 3 plots both $-\Pi^H$ and $-\Pi^{NH}$ against τ_0 for the complete set of outcomes ($2^{10} = 1024$ paths). The values are per \$100 initial premium.

Figure 2: Probability mass function of τ : different volatilitiesFigure 3: Hedging and no-hedging losses, with $r = 5\%$ and $g = 10\%$

The dynamic delta hedging strategy results in no losses. Without hedging, the range of potential losses by each random trigger time has a decreasing trend because a later trigger time implies additional periods of fee revenue and fewer periods of any rider guarantee payout. The effect of the volatility σ is particularly visible for those pathwise outcomes where $\tau = \infty$. When $\sigma = 15\%$ there is an 87% probability of a positive terminal account value but the gains are small. On the other hand, there is only a 50% probability that $\tau = \infty$ when $\sigma = 30\%$ but the potential profits are large due to the high fees. Figure 4 shows the cumulative distribution function of the profits when there is no hedging.

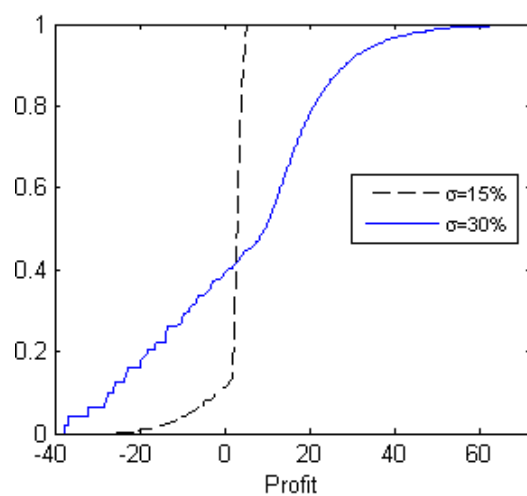
We present several risk measures for the no-hedging profit Π^{NH} under \mathbb{P} . The standard deviation is denoted $SD(\Pi)$. The tail value at risk is $TVaR_\gamma(\Pi) := E_{\mathbb{P}}[-\Pi | \Pi \leq -VaR_\gamma(\Pi)]$ where $VaR_\gamma(\Pi) = -\inf\{x : \mathbb{P}(\Pi \leq x) > \gamma\}$. Table 6 shows the values for this sensitivity analysis of σ . Using the real world probability measure only amplifies the effect of σ on the insurer's risk and highlights the importance of a thorough hedging scheme.

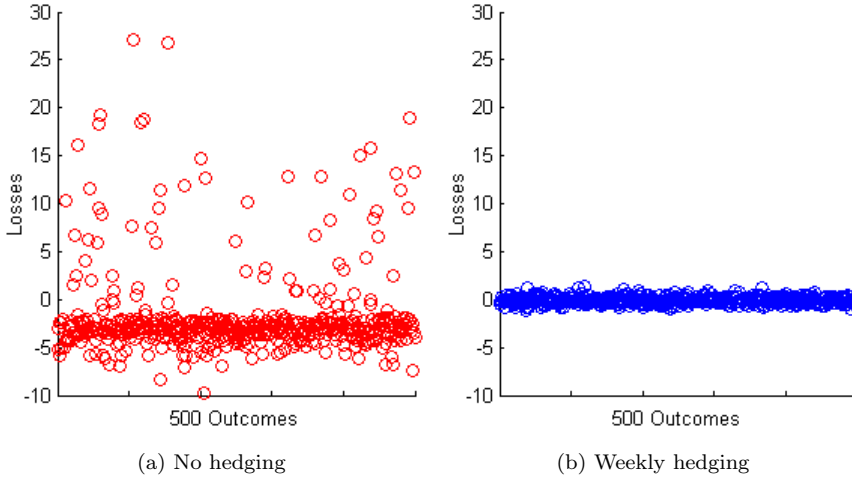
5.3.1 Hedging in a Continuous Model

In the binomial model a perfect hedge is attainable. Suppose instead the underlying asset follows the geometric Brownian motion process given by (4). A perfect hedge in this case entails continuously re-balancing the hedging positions by taking a position at any time t of $\frac{W}{S} \frac{\partial U}{\partial W}$ units of S (see 6). In practice, the positions will be rebalanced only a finite number of times each year which introduces hedging errors. We model the fees and withdrawals to occur only at year-end in order to contrast

Values per \$100	$\sigma = 15\%$	$\sigma = 30\%$
$E_{\mathbb{P}}(\Pi^{NH})$	1.84	4.19
$SD_{\mathbb{P}}(\Pi^{NH})$	4.28	21.34
$TVaR_{0.10}(\Pi^{NH})$	9.30	32.60

Table 6: Profit metrics for no hedging (no lapses)

Figure 4: CDF of Π^{NH} with respect to \mathbb{P}

Figure 5: Continuous model with $g=10\%$, $r=5\%$, $\mu=7.5\%$, $\sigma=15\%$, $\alpha=45\text{bps}$

Values per \$100	No Hedging	Hedging (Weekly)
$E_{\mathbb{P}}[\Pi]$	1.86	0.07
$SD_{\mathbb{P}}[\Pi]$	4.63	0.36
$TVaR_{0.10}(\Pi)$	10.15	0.61

Table 7: Profit metrics for continuous model with weekly hedging and no hedging

with the previous result in the binomial model for $\delta t = 1$. This differs from the continuous model of Hyndman and Wenger [15] where fees and withdrawals are deducted continuously.

The parameters used are $P = 100$, $g = 10\%$, $r = 5\%$, $\mu = 7.5\%$, $\sigma = 15\%$, and $T = 10$. We used Monte Carlo simulation to obtain $\alpha^* \approx 45\text{bps}$ (50,000 paths were simulated). We analyzed the effectiveness of a dynamic hedging strategy with weekly re-balancing for 500 path outcomes generated under \mathbb{P} . For $t \in \{0, \frac{1}{52}, \frac{2}{52}, \dots, \frac{519}{52}, 10\}$ and $w \in \mathbb{R}_+$, Monte Carlo simulations (using 1000 paths) yielded $U_t(w-1)$ and $U_t(w+1)$. We approximated $\frac{\partial U}{\partial W}$ with $\Delta_t(W_t) = (U_t(W_t + 1) - U_t(W_t - 1))/2$ where the same set of generated paths was used to obtain both values in the numerator. Using the same paths and taking the central difference has been shown to reduce variability of results (13). Figure 5 displays the discounted losses for no hedging and for weekly hedging for each generated path. Based on the simulations, $\mathbb{P}(\tau = \infty) = 84.4\%$. As supported by Table 7, the weekly hedging considerably mitigates the equity risk. In contrast to the case when the underlying model is binomial, negative hedging errors arise when the underlying model is continuous.

5.4 The Fair Rider Fee with Surrenders

We next compare our results for α^* when early surrenders are permitted with those in the literature. For the parameter set of $g = 7\%$, $r = 5\%$, and $k_i = 1\%$ for all i , Table 8 compares the binomial model with $\delta t = 1$ to Milevsky and Salisbury [19]. Although the results are proportionally closer, as compared to Table 1, it is inconclusive if the differences are mostly due to $\delta t = 1$ or if the results presented by Milevsky and Salisbury [19] in the lapse case suffer from the same inaccuracies as in the no-lapse case.

We apply the Asian approximation method with the parameters $g = 10\%$, $r = 5\%$, $\sigma = 20\%$,

$\sigma(\%)$	15	18	20	25	30
Milevsky and Salisbury [19]	97	136	160	320	565
Binomial ($\delta t = 1$)	33	89	138	283	455

Table 8: Comparison of α^* to previous results; with $g = 7\%$, $r = 5\%$, and $k = 1\%$.

n	$\alpha^*(\text{bps})$	$V_0(\alpha=146.4)(\$)$	$\alpha^*(\text{actual})$
1	131.00	99.689	130.54
2	141.98	99.933	141.75
3	143.37	99.949	
4	146.04	99.994	
5	146.40	100	
6	146.70	100.005	

Table 9: Asian approximation results - lapses

and $k = 3\%$ in Table 9. The convergence is slower than in the no-lapse case, but that is a result of the early surrender decisions which are being approximated. This is consistent with the findings of Costabile et al. [7]. The rightmost column shows α^* under the original binomial model. The increase in α^* when n is increased from one to two suggests that a sizable portion of the differences in Table 8 can be attributed to the low value of n in the binomial model.

We set r equal to the instantaneous risk-free rate long term mean and σ equal to the variance long term mean used in the stochastic interest rate and volatility processes in Bacinello et al. [2]. We found that comparing V_0 for varying α , in the no lapse case the binomial model provides close estimates even for $\delta t = 0.5$. In Table 10 we list the difference in the contract value between the two methods for varying α and $P = 100$, $g = 10\%$, $r = 3\%$, $\sigma = 20\%$, and $k = 3\%$. The models have fundamental differences and we do not expect to attain exact results in the limit.

Sensitivity results for g , r , and σ are shown in Table 11. The baseline case is set to $g = 10\%$, $r = 5\%$, $\sigma = 20\%$, and a CDSC of $k = 3\%$. The fair fee α^* is increasing with g and σ but decreasing with r , however, the fair fee is most sensitive to r . The sensitivity of the fair fee to r is due to the long duration of the contract. Therefore, incorporating a stochastic interest rate model is justified, though beyond the scope of this paper.

$\alpha(\%)$	1	2	3	4	5
$V_0^B(\alpha) - V_0^{BMOP}(\alpha)^{a,b}$: (no lapse)	-0.186	-0.113	-0.035	0.05	0.096
$V_0^B(\alpha) - V_0^{BMOP}(\alpha)$: (lapse)	0.153	0.546	0.75	0.78	1.04

^a V_0^B refers to the binomial method, with $\delta t = 0.5$.

^b V_0^{BMOP} refers to Bacinello et al. [2].

Table 10: Comparison of V_0 with previous results: $g = 10\%$, $P = 100$, $r = 3\%$, $\sigma = 20\%$, and $k = 3\%$.

$g\%$	α^* (bps)	$V_0(\alpha_1)$ (\$)	$\sigma\%$	α^*	$V_0(\alpha_1)$	$r\%$	α^*	$V_0(\alpha_1)$
5	30	97.21	10	10	97	1	1199	108.21
6	47	97.87	15	44	97.84	2	673	105.54
7	68	98.44	18	87	99.08	3	397	103.29
8	90	98.95	20	142	100	4	244	101.43
9	110	99.38	25	318	102.46	5	142	100
10	142	100	30	562	105.12	6	77	98.87

^a Baseline case is $g = 10\%$, $r = 5\%$, $\sigma = 20\%$, $k = 3\%$, $\alpha_1 = 142\text{bps}$.

^b For the first column, $\delta t = 1$ for $g \leq 9\%$. All other values use $\delta t = 2$.

Table 11: Sensitivity results for α^*

Under the parameters $g = 10\%$, $r = 5\%$, $\sigma = 25\%$, and $\delta t = 1$, the impact of the CDSC schedule on α^* is shown in Table 12. Allowing surrenders with no penalties, the fair fee will be exorbitant to compensate for this option. As the penalties increase, the fee approaches the corresponding fee in the no-lapse model. For sufficiently high penalties, the option to surrender yields no marginal value.

5.5 Hedging and No Hedging with Surrenders

We consider the parameters: $P = 100$, $g = 10\%$, $r = 5\%$, $\sigma = 25\%$, and $\delta t = 1$. The drift of S is $\mu = 7.5\%$. The surrender charge schedule applied is $k_i = \max(.09 - .01i, 0)$ for $i = 1 \dots 10$. Figure 6 plots the aggregate losses, discounted to time zero, for the set of all outcomes for both the no-surrender model and the model with early surrenders. The respective fair fees are charged. In Figure 6b the no-hedging results are denoted by L and T: the former are outcomes where it is optimal to lapse while the latter are those for which no lapse occurs.

Table 13 shows the \mathbb{P} -distribution of trigger times and surrender times, where η^* denotes an optimal early surrender. Note that $\mathbb{P}(\tau = \infty) \approx 60\%$ when surrenders are not allowed, but this reduces to $\mathbb{P}(\tau = \infty) \approx 0.65\%$ when surrenders are permitted. Allowing lapses causes a shift as it becomes preferable in many outcomes when the market is doing well for the policyholder to lapse rather than face the likelihood of the rider maturing without being triggered.

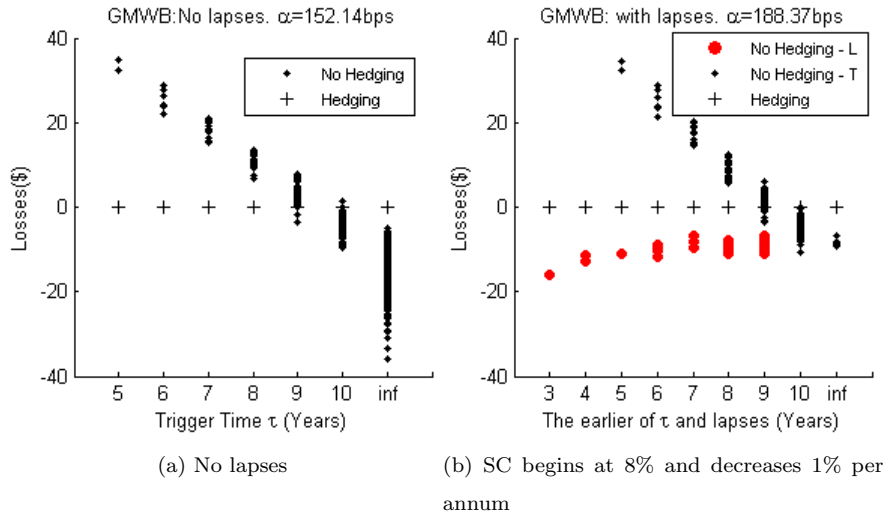
For the outcomes where it is optimal to lapse, the profits to the insurer are decreasing for years 3 to 7. This is due to the design of the surrender charge schedule k_i . The higher surrender charge in earlier years outweighs the additional fees received when lapses occur later.

Numerical results for the value of the option to surrender, L_0 , are presented in Figure 7. When α is small, there is little incentive to surrender early and $L_0 \approx 0$. For larger values of α there is incentive to surrender and avoid paying future fees. This relationship is reflected in the growth of L_0 with α .

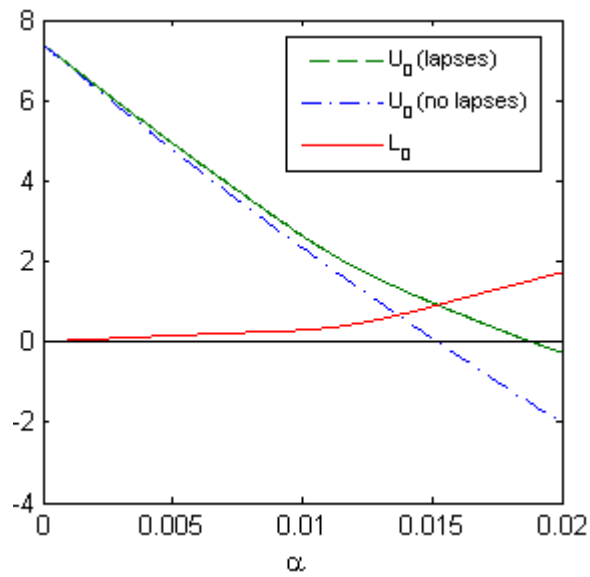
6 Extending the Model: Including Mortality Risk

The simplification of disregarding mortality was used in several papers for GMWBs including Milevsky and Salisbury [19] and Dai et al. [9]. Mortality factors do need to be considered in practice. Depending on the goal of the analysis, the level of precision attained by including mortality may not justify the added complexity and dimensionality of the model. In particular, in the papers mentioned the focus was on studying the optimal policyholder behaviour strategy and including mortality only detracts from the presentation of the results.

Description of Schedule	α^* (bps)
No-Lapse Model	152
$k_i = 0$ for $i = 1, \dots, 9$	491
$k_i = 1\%$ for $i = 1, \dots, 9$	430
$k_i = 3\%$ for $i = 1, \dots, 9$	309
$k_i = 5\%$ for $i = 1, \dots, 9$	217
$k_i = 7\%$ for $i = 1, \dots, 9$	169
$k_i = 8\%$ for $i = 1, \dots, 9$	155
$k_i \geq 8.38\%$ for $i = 1, \dots, 9$	152
$k_i = (10 - i)\%$, for $i = 1, \dots, 9$	171
$k_i = (9 - i)\%$, for $i = 1, \dots, 9$	188

Table 12: Impact of k on α^* Figure 6: Hedging and no hedging, with and without lapses: $g = 10\%$, $r = 5\%$, $\sigma = 25\%$.

	No Lapses	Model with Lapses	
i	$\mathbb{P}(\tau = i)$	$\mathbb{P}(\tau = i)$	$\mathbb{P}(\eta^* = i)$
3	0	0%	20.28%
4	0	0%	16.73%
5	2.90%	2.90%	4.91%
6	5.80%	5.80%	8.11%
7	7.83%	7.83%	3.57%
8	6.29%	9.08%	4.42%
9	9.48%	7.37%	2.37%
10	7.23%	5.98%	0
∞	60.47%	.65%	0
Sum	1	39.61%	60.39%

Table 13: Probability distribution of τ and lapses for Figure 6Figure 7: Value of L_0 : $g = 10\%$, $r = 5\%$, $\sigma = 25\%$, $\delta t = 1$, and a declining SC schedule

Mortality risk is typically assumed to be independent of financial risk. Further, under the assumption of independent lives and deterministic forces of mortality (hazard rates) a simple application of the strong law of large numbers justifies the claim that mortality risk is diversifiable. By issuing a sufficiently large portfolio of homogeneous policies the insurer can completely account for the mortality risk by taking the expected value of claim payments under the appropriate mortality probability distribution (5). Therefore under these assumptions mortality risk is not priced by capital markets in an economic equilibrium (no-arbitrage) approach and there is no difference between the physical and risk-neutral measures (20). In a stochastic mortality framework the non-diversifiable component of mortality risk must be priced into the contract.

Milevsky et al. [20] list capacity constraints in immediate annuity markets as one of several industry trends which justify charging for mortality risk. We remark that in variable annuity markets, both finite demand and regulatory limits on capital at risk lend support to modeling capacity constraints in order to determine whether there is a non-negligible impact.

The effect of mortality for GMWBs clearly depends on the death benefits (DBs). When benefit payments are similar for both death and survival, there is minimal impact. Indeed, Bacinello et al. [2] found that guaranteed minimum death benefit (GMDB) riders add little value to the contract in the presence of other living benefit riders and a relatively short maturity.

We extend the model from Section 2 and Section 3 to include mortality under the independence of lives assumption and deterministic forces of mortality. It is straightforward to obtain the price processes V and U , which for each insured are dependent on the survival status. The rider fee is obtained assuming diversifiable mortality risk, as is the hedging portfolio; however, we consider a numerical simulation to emphasize that under capacity constraints and finite number of policies there is mortality risk and the product is not fully hedged.

6.1 Mortality Risk Framework

In this section we establish a mortality framework. The classical actuarial theory and notation used follows that of Bowers et al. [4]. In addition, the measure-theoretic aspects and inclusion of counting processes follows closely the frameworks of Møller [21] and Wang [25].

Assumption 12. *Homogeneous policies are issued to a pool of l_x policyholders, each of age x . Measured from issue date, the random times of death, denoted by $\{T_j^x; j = 1, \dots, l_x\}$ where T_j^x is the time of death for policyholder j , are absolutely continuous, independent and identically distributed, and lie on a probability space $(\Omega^M, \mathcal{F}^M, \mathbb{P}^M)$.*

Consider a representative random variable T^x where T^x has the same distribution as T_j^x . The support of T^x is $[0, T^*)$ where $T^* \leq \infty$ is the maximum remaining lifetime for a person age x . Corresponding to the binomial model with $\delta t = 1/n$ and $n \in \mathbb{N}_+$, let K^x denote the period in which death occurs. Then $K^x = \lceil T^x / \delta t \rceil$. In other words, $K^x = i$ is equivalent to $(i-1)\delta t < T^x \leq i\delta t$. For $j = 1, \dots, l_x$, define the counting processes

$$D^{x,j} = \{D_i^{x,j} := \mathbf{1}_{\{K_j^x \leq i\}}; i = 1, \dots, N\}.$$

We work with the filtration generated by $\{D^{x,j}\}_{1 \leq j \leq l_x}$. The filtration is $\mathbb{F}^{M,x} := \{\mathcal{F}_i^{M,\{x,l_x\}}\}_{1 \leq i \leq N}$ where $\mathcal{F}_i^{M,\{x,l_x\}} := \mathcal{F}_i^{M,x,1} \vee \dots \vee \mathcal{F}_i^{M,x,l_x}$ and $\mathcal{F}_i^{M,x,j} = \sigma(D_l^{x,j}; l = 1, \dots, i)$. We work with the resulting filtered probability space $(\Omega^M, \mathcal{F}_N^{M,\{x,l_x\}}, \mathbb{F}^{M,x}, \mathbb{P}^M)$.

Remark 6. *The notation $\mathcal{G} \vee \mathcal{H}$, where \mathcal{G} and \mathcal{H} are σ -algebras, means the σ -algebra generated by $\mathcal{G} \cup \mathcal{H}$.*

We define the process which produces 1 while the insured j is still alive by $A_i^{x,j} := 1 - D_i^{x,j}$ for $i \in \mathcal{I}_N$.

By Assumption 12, T^x has a density function f_{T^x} . Its cumulative distribution function is denoted $F_{T^x}(t) := \mathbb{P}(T^x \leq t)$. The deterministic force of mortality, $\mu_x(t)$, is defined as the conditional probability density function of T^x at time t , given survival to that time. Then

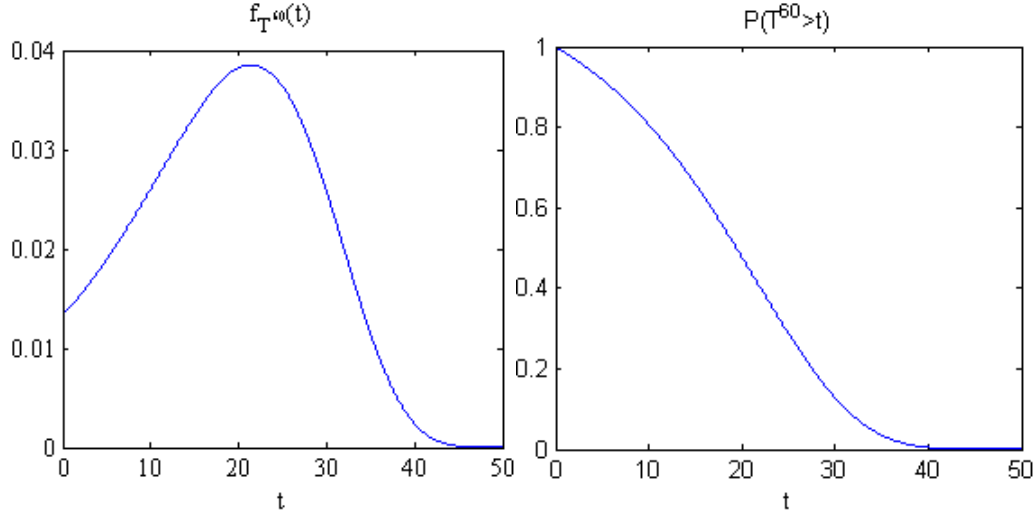


Figure 8: Using the Makeham law, with $A = 0.7 \times 10^{-3}$, $B = 0.05 \times 10^{-3}$ and $c = 10^{0.04}$

$$\mu_x(t) := \frac{f_{T^x}(t)}{1 - F_{T^x}(t)}. \quad (33)$$

We introduce some additional actuarial notation:

$$\begin{aligned} {}_j p_{x+i} &:= \mathbb{P}(K^x > i + j | K^x > i) = \mathbb{P}(T^x > (i + j)\delta t | T^x > i\delta t), \\ {}_{j|l} q_{x+i} &:= \mathbb{P}(i + j < K^x \leq i + j + l | K^x > i), \end{aligned}$$

and we write p_{x+i} for ${}_1 p_{x+i}$, ${}_j q_{x+i}$ for ${}_{0|j} q_{x+i}$, and q_{x+i} for ${}_1 q_{x+i}$. It follows that

$${}_i q_x = F_{T^x}(i\delta t), \quad {}_i p_x = 1 - {}_i q_x, \quad \text{and} \quad {}_{j|l} q_{x+i} = {}_{j+l} q_{x+i} - {}_j q_{x+i}.$$

From (33) we have $f_{T^x}(i\delta t) = \mu_x(i\delta t) {}_{i\delta t} p_x$ and ${}_j p_{x+i} = e^{-\int_0^{j\delta t} \mu_{x+i\delta t}(u) du}$ (see Bowers et al. [4] for details). Note that F_{T^x} , f_{T^x} , and μ_x are defined on the reals, while ${}_j p_{x+i}$ and ${}_{j|l} q_{x+i}$ are defined on the integers. Bowers et al. [4] provide several analytical laws of mortality.

Definition 13. *Under the Makeham law*

$$\mu_x(t) := A + Bc^{x+t}$$

where $B > 0$, $A \geq -B$, $c > 1$ and $x + t \geq 0$.

As a result, under the Makeham law:

$${}_i p_x = \exp\left(-i\delta t A - \frac{B}{\ln(c)}(c^{x+i\delta t} - c^x)\right).$$

Example 1. *The parameters used to develop the illustrative life table under the Makeham law in Bowers et al. [4] are: $A = 0.7 \times 10^{-3}$, $B = 0.05 \times 10^{-3}$ and $c = 10^{0.04}$. Figure 8 plots both $f_{T^x}(t)$ and $\mathbb{P}(T^x > t)$ for $x = 60$ and $t \in [0, 50]$.*

We state one additional useful result from Wang [25]. For $i \leq j$,

$$\mathbb{P}(T^x > j\delta t | \mathcal{F}_i^{M,x}) = (1 - D_i^x)_{j-i} p_{x+i}$$

and

$$\mathbb{P}(i\delta t < T^x \leq j\delta t | \mathcal{F}_i^{M,x}) = (1 - D_i^x)_{j-i} q_{x+i}.$$

6.2 Death Benefit Design

We consider both the ratchet DB and the return of premium DB. The ratchet DB has the feature that on each ratchet date, the death benefit base will increase to the current account value, provided the account value is higher. Let

$$0 \leq t_1 < t_2 < \dots < t_m \leq T$$

represent the set of ratchet dates prior to maturity. Then the rescaled set, in terms of binomial time periods, is

$$I = \left\{ \frac{t_1}{\delta t}, \frac{t_2}{\delta t}, \dots, \frac{t_m}{\delta t} \right\} \subset \mathcal{I}_N.$$

The GMWB and GMDB are treated as one rider with the aim of solving for the fair fee α^* as before. Alternatively, one could separate the two and specify the GMDB rider fee exogenously. Let DB_i be the death benefit guarantee base at time point i , with $DB_0 = P$. Then $DB_i = db(i, W_{i-}, DB_{i-1})$, where $db : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined as

$$\begin{cases} db(0, x, y) = x, \\ db(i, x, y) = \max \left(w(x) \mathbf{1}_{\{i \in I\}}, \frac{w(x)}{xe^{-\alpha}} y \right). \end{cases} \quad (34)$$

If $I = \emptyset$, then the ratchet DB reduces to a simple return of premium DB.

Note that $DB_i = 0$ for $i \geq \tau$. However we assume that conditional on survival to the trigger date, the guaranteed payments are paid regardless of life status; that is, the present value of the remaining payments is paid upon death if trigger has previously occurred. The death benefit of $\max(DB_i, W_{i+1-})$ is paid at time $(i+1)\delta t$, if death occurs during the $(i+1)$ th period but prior to trigger time. In the limit as $\delta t \rightarrow 0$ this corresponds to the death benefit being paid at the instantaneous time of death.

The death benefit base in (34) is reduced by withdrawals in a pro-rata manner, meaning it is reduced by the same proportion as the account value. Another method is called dollar-for-dollar withdrawal adjusted. Assume a policyholder holds a deep in the money GMDB, with $DB_i \gg W_i$ (where $x \gg y$ means y is much less than x). By withdrawing $0.9W_i$ and ignoring surrender charges, under the dollar-for-dollar reduction method the policyholder holds a GMDB with only 10% of the previous account value but a death benefit base of $DB_i - 0.9W_i \gg 0$. Under the pro-rata method, the new death benefit base is $0.1DB_i \ll DB_i - 0.9W_i$.

6.3 Pricing and Hedging a Single Contract

A key underlying assumption for the remainder of our work is stated.

Assumption 14. *There is independence between biometric and financial risks. Let $(\Omega^S, \mathcal{F}_N^S, \mathbb{F}^S, \mathbb{Q}^S)$ and $(\Omega^M, \mathcal{F}_N^M, \mathbb{F}^{M,x}, \mathbb{P}^M)$ be the filtered probability spaces constructed in Section 2 and Section 6.1 respectively. We work with the product space $(\Omega, \mathcal{F}_N, \mathbb{F}, \mathbb{Q})$ where $\Omega := \Omega^M \times \Omega^S$, $\mathbb{F} := \{\mathcal{F}_i\}_{i=0}^N$, $\mathcal{F}_i := \mathcal{F}_i^{M, \{x, l_x\}} \times \mathcal{F}_i^S := \sigma(\{A \times B : A \in \mathcal{F}_i^{M, \{x, l_x\}}, B \in \mathcal{F}_i^S\})$ and $\mathbb{Q} := \mathbb{P}^M \times \mathbb{Q}^S$.*

We present the more general model allowing for early surrenders and as in Section 3 optimal policyholder behaviour is assumed. The no-lapse model is obtained under the following assumption.

Assumption 15. (No-lapse model) The surrender charges satisfy $k_i = 1$ for all $i < N$ and $k_N = 0$.

This implies that the set of admissible lapse strategies is $\mathbb{L}_0 = \{N\}$.

Without loss of generality, from now until after Subsection 6.4 we consider the case of a single contract sold to an individual at age x , that is we let $l_x = 1$. The value process $\{V_i^M\}_{0 \leq i \leq N}$ is defined as

$$V_i^M = A_i^x \max_{\eta \in \mathbb{L}_i, \bar{\tau}_i} E_{\mathbb{Q}} \left[D_{\bar{\tau}_i \wedge \eta}^x \left(\max(DB_{K^x-1}, W_{K^x-}) e^{-\bar{r}(K^x-i)} + Ga_{\overline{K^x-1-i}} \right) \right. \\ \left. + A_{\bar{\tau}_i \wedge \eta}^x \left(Ga_{\overline{\eta-i}} + W_{\eta}(1 - k_{\eta}) e^{-\bar{r}(\eta-i)} \right) \middle| \mathcal{F}_i \right].$$

Observe that all $\eta \in \mathbb{L}_i$ are \mathbb{F}^S -stopping times and are independent of the mortality probability measure. Any lapse strategy η is only exercised if the insured is still alive. It remains true that the optimal lapse strategy must lie in $\mathbb{L}_i, \bar{\tau}_i \subset \mathbb{L}_i$.

Conditioning on the time of death and taking the expectation with respect to \mathbb{P}^M (justified by the independence of \mathbb{Q}^S and \mathbb{P}^M) we obtain

$$V_i^M = A_i^x V_i,$$

where

$$V_i = \max_{\eta \in \mathbb{L}_i, \bar{\tau}_i} V_i^{\eta}$$

and

$$V_i^{\eta} = E_{\mathbb{Q}^S} \left[\sum_{j=i}^{\bar{\tau}_i \wedge \eta-1} j-i | q_{x+i} \left(\max(DB_j, W_{j+1-}) e^{-\bar{r}(j+1-i)} + Ga_{\overline{j-i}} \right) \right. \\ \left. + \bar{\tau}_i \wedge \eta - i p_{x+i} \left(Ga_{\overline{\eta-i}} + W_{\eta}(1 - k_{\eta}) e^{-\bar{r}(\eta-i)} \right) \middle| \mathcal{F}_i^S \right]. \quad (35)$$

The definition for the fair fee rate α^* remains unchanged and it satisfies $V_0^M = P$. Select any $\eta \in \mathbb{L}_0$. Denote ${}^R\tilde{V}_i^{\eta}$ to be the total contract payouts up to time point i under this surrender strategy and discounted to $t = 0$. Then

$${}^R\tilde{V}_i^{\eta} = \sum_{j=0}^{\tau \wedge \eta \wedge i-1} (A_j^x - A_{j+1}^x) \left[\max(DB_j, W_{j+1-}) e^{-\bar{r}(j+1)} + Ga_{\overline{j}} \right] + A_{\tau \wedge \eta \wedge i}^x Ga_{\overline{\eta \wedge i}}.$$

Let ${}^RV_i^{\eta} := E_{\mathbb{P}^M}[{}^R\tilde{V}_i^{\eta}]$. Then we have

$${}^RV_i^{\eta} = \sum_{j=0}^{\tau \wedge \eta \wedge i-1} j | q_x \left[\max(DB_j, W_{j+1-}) e^{-\bar{r}(j+1)} + Ga_{\overline{j}} \right] + \tau \wedge \eta \wedge i p_x Ga_{\overline{\eta \wedge i}}.$$

For any $0 \leq i \leq N$, define a rescaled filtration $\mathbb{F}^{S,i} = \{\mathcal{F}_j^{S,i} := \mathcal{F}_{j+i}^S; 0 \leq j \leq N-i\}$. Then the process

$$Y^{\eta,i} = \left\{ Y_j^{\eta,i} = e^{-\bar{r}((j+i) \wedge \eta)} ({}_{(j+i) \wedge \eta} p_x V_{(j+i) \wedge \eta}^{\eta} + {}^RV_{(j+i) \wedge \eta}^{\eta}) \right\}_{0 \leq j \leq N-i} \quad (36)$$

is a $(\mathbb{Q}^S, \mathbb{F}^{S,i})$ martingale. The optimal surrender strategy, $\hat{\eta}_i$, is given by (21) (the proof is similar and uses the martingale (36)).

Since $\{W_i, DB_i\}_{i=0,1,\dots,N}$ is a 2-dimensional Markov process we have

$$V_i^M = A_i^x v(i, W_i, DB_i),$$

where $v : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is recursively defined by

$$v(N, x, y) = x$$

and for $0 \leq i \leq N-1$

$$\begin{aligned} v(i, x, y) = & \max\{e^{-\bar{r}}[p_{x+i}(G + pv(i+1, w(ux), db(i+1, ux, y)) \\ & + qv(i+1, w(dx), db(i+1, dx, y))) \\ & + q_{x+i}((p \max(y, ux) + q \max(y, dx))\mathbf{1}_{\{x>0\}} + \mathbf{1}_{\{x=0\}}Ga_{\overline{N-i}})], x(1-k_i)\}. \end{aligned}$$

This implies the boundary condition $v(i, 0, y) = Ga_{\overline{N-i}}$.

The rider value process must account for the following cash flow components. The rider fee is paid prior to trigger while the insured is alive and has not surrendered. If surrender occurs prior to trigger time then no cost is incurred for the GMWB rider. In the event that no surrender occurs and the insured is alive at trigger time, the periodic GMWB guarantee is paid out until maturity regardless of death. If death occurs prior to the earlier of trigger time or surrender time, then any excess of the death benefit over the current account value is a cost incurred by the rider. Putting this together, we have

$$\begin{aligned} U_i^M = & A_i^x \max_{\eta \in \mathbb{L}_i, \bar{\tau}_i} E_{\mathbb{Q}} \left[\sum_{j=i+1}^{\eta} e^{-\bar{r}(j-i)} \left[A_{\bar{\tau}_i}^x (G - W_{j-} e^{-\bar{\alpha}})^+ - A_j^x W_{j-} (1 - e^{-\bar{\alpha}}) \right. \right. \\ & \left. \left. - k_{\eta} W_{\eta} e^{-\bar{r}(\eta-i)} A_{\eta}^x \right] + D_{\eta}^x (DB_{K^x-1} - W_{K^x-})^+ e^{-\bar{r}(K^x-i)} | \mathcal{F}_i \right]. \end{aligned} \quad (37)$$

Then $U_i^M = A_i^x U_i = A_i^x u(i, W_i, DB_i)$, where $u : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is described by

$$\begin{cases} u(N, x, y) = 0, \\ u(i, x, y) = \max\{e^{-\bar{r}}(pu^-(i+1, ux, y) + qu^-(i+1, dx, y)), -k_i x\}, \end{cases} \quad (38)$$

and $u^- : \mathcal{I}_N^+ \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is given by

$$\begin{aligned} u^-(i, 0, y) &= G\ddot{a}_{\overline{N-i+1}}, \\ u^-(i, x, y) &= p_{x+i-1}[(G - xe^{-\bar{\alpha}})^+ - x(1 - e^{-\bar{\alpha}}) + u(i, w(x), db(i, x, y))] \\ &\quad + q_{x+i-1}(y - x)^+. \end{aligned} \quad (39)$$

The notation $\ddot{a}_{\overline{i+1}} = 1 + a_{\overline{1}}$ is an annuity due. Under Assumption 15 it is easy to check that the term $(-k_i x)$ is never binding. Note that $A_{i-1}^x u^-(i, W_{i-}, DB_{i-1})$ is $\mathcal{F}_i^S \times \mathcal{F}_{i-1}^{M, \{x, t_x\}}$ -measurable. It is the rider value at time point i evaluated once the market movement for the past period is known, but prior to any transactions occurring (i.e. fees, withdrawals or death benefits). That is, the insurer knows the exact market growth in the funds over the past period but is waiting to find out about the status of the policyholder.

We denote $\{U_i^{M, NL}\}$ to refer to (37) when Assumption 15 is in place. The marginal rider value from the option to surrender is $L_i^M := U_i^M - U_i^{M, NL} \geq 0$ and can be written as

$$\begin{aligned} L_i^M = & A_i^x \max_{\eta \in \mathbb{L}_i, \bar{\tau}_i} E_{\mathbb{Q}} \left[\sum_{j=\eta+1}^N e^{-\bar{r}(j-i)} \left[A_j^x W_{j-} (1 - e^{-\bar{\alpha}}) - A_{\bar{\tau}_i}^x (G - W_{j-} e^{-\bar{\alpha}})^+ \right. \right. \\ & \left. \left. - A_{\eta}^x \left[k_{\eta} W_{\eta} e^{-\bar{r}(\eta-i)} + D_{\eta}^x (DB_{K^x-1} - W_{K^x-})^+ e^{-\bar{r}(K^x-i)} \right] | \mathcal{F}_i \right] \right]. \end{aligned} \quad (40)$$

Then $L_i^M = A_i^x l(i, W_i, DB_i)$, where $l : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$\begin{aligned} l(N, x, y) &= 0, \\ l(i, x, y) &= \max\{p_{x+i}e^{-\bar{r}}(pl(i+1, w(ux), db(i+1, ux, y)) \\ &\quad + ql(i+1, w(dx), db(i+1, dx, y))), -u^{NL}(i, x, y) - k_i x\}. \end{aligned}$$

Backward induction verifies that $l(i, x, y) = u(i, x, y) - u^{NL}(i, x, y)$.

Proposition 16. *For any $\alpha > 0$ we have*

$$V_i^M = U_i^M + A_i^x W_i \quad (41)$$

or equivalently

$$V_i^M = U_i^{M,NL} + L_i^M + A_i^x W_i \quad (42)$$

\mathbb{Q} -a.s. for all $0 \leq i \leq N$.

Proof. The equality (41) can be proved either directly from (35) and (37) or through backward induction applied to the functions v , u , and u^- . The procedure is similar to the proof of Theorem 7. We omit the details. \square

The \mathbb{F}^s -adapted portfolio process $\{\Delta_i\}$ is defined by $\Delta_i = \Delta(i, S_i, W_i, DB_i)$, where $\Delta : \mathcal{I}_{N-1} \times \mathbb{R}_+^3 \mapsto \mathbb{R}$ is given by

$$\Delta(i, w, x, y) = \frac{u^-(i+1, ux, y) - u^-(i+1, dx, y)}{wu - wd}. \quad (43)$$

Note that $\Delta(i, w, 0, y) = 0$. For a given policy, the insurer follows $\{\Delta_i\}$ only up until the death of the policyholder or the surrender of the policy.

Similar to Section 3, we define a consumption process $\{C_i\}_{0 \leq i \leq N-1}$ where $C_i = c(i, W_i, DB_i)$ and $c : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined as

$$\begin{aligned} c(i, x, y) &:= v(i, x, y) - e^{-\bar{r}} [p_{x+i}(G + pv(i+1, w(ux), db(i+1, ux, y)) \\ &\quad + qv(i+1, w(dx), db(i+1, dx, y))) \\ &\quad + q_{x+i}((p \max(y, ux) + q \max(y, dx)) \mathbf{1}_{\{x>0\}} + \mathbf{1}_{\{x=0\}} Ga_{N-i})] \\ &= u(i, x, y) - e^{-\bar{r}} [pu^-(i+1, ux, y) + qu^-(i+1, dx, y)]. \end{aligned} \quad (44)$$

The second equality can be verified using Proposition 16, similar to (28). Under Assumption 15 we have $C \equiv 0$.

Construct the replicating portfolio by starting with initial capital $X_0 = x_0$ and following the portfolio process $\{\Delta_i\}$. For $i \in \mathcal{I}_N^+$ we have

$$\begin{aligned} X_i &= (X_{i-1} - A_{i-1}^x (\Delta_{i-1} S_{i-1} + C_{i-1})) e^{\bar{r}} + A_{i-1}^x \Delta_{i-1} S_i + A_i^x [F_i - (G - W_{i-} e^{-\bar{\alpha}})^+] \\ &\quad - (A_{i-1}^x - A_i^x) [(DB_{i-1} - W_{i-})^+ \mathbf{1}_{\{\tau \geq i\}} + G \ddot{a}_{N-i+1} \mathbf{1}_{\{\tau < i\}}]. \end{aligned} \quad (45)$$

The fees, payouts, portfolio process, and consumption process have all been defined in \mathbb{F}^s . Of course they are only applicable while the policy is in force (prior to death or surrender). For that reason, the terms are accompanied by A_i^x factors in (45). Given a surrender strategy $\eta \in \mathbb{L}_0$, the insurer will close out its position at time point η and the process of interest is $\{X_{i \wedge \eta}\}_{0 \leq i \leq N}$. The time zero profit is $\Pi = e^{-\bar{r}\eta} X_\eta$, since if death occurs prior to η then the portfolio remains unchanged for all periods between death and η , aside from interest accumulation.

Although we no longer have almost sure equivalence of U^M and X with respect to the product measure \mathbb{Q} , an analogous result holds by considering the conditional expectation with respect to \mathbb{P}^M .

Theorem 17. *Suppose the fee rate α is charged and the initial capital is $x_0 = U_0^M$. Then the following relation holds between X_i , described by (45), and U_i^M , given by (37):*

$$\mathbb{Q}^S(\mathbb{E}_{\mathbb{P}^M}[X_i - U_i^M] = 0) = 1$$

for all $i \in \mathcal{I}_N$.

Proof. See Appendix A □

6.4 Diversification of Mortality Risk for Multiple Contracts

Suppose homogeneous policies are sold to l_x independent policyholders aged x , each with an initial premium of P and the fair rider fee α^* is charged. For the pool of l_x insureds, the number of deaths between time $i\delta t$ and $(i+1)\delta t$ is

$$\mathcal{D}_i^{l_x, x} := \sum_{j=1}^{l_x} (A_i^{x, j} - A_{i+1}^{x, j})$$

for $i \in \mathcal{I}_{N-1}$. The number of members alive at time i is

$$\mathcal{A}_i^{l_x, x} = \sum_{j=1}^{l_x} A_i^{x, j} = l_x - \sum_{j=1}^{i-1} \mathcal{D}_j^{l_x, x}.$$

By the strong law of large numbers (SLLN), as $l_x \rightarrow \infty$,

$$\frac{\mathcal{D}_i^{l_x, x}}{l_x} \rightarrow {}_i q_x \quad \text{and} \quad \frac{\mathcal{A}_i^{l_x, x}}{l_x} \rightarrow {}_i p_x$$

\mathbb{P}^M -a.s., for all $i \in \mathcal{I}_N$.

The aggregate replicating portfolio process is the sum of the individual replicating portfolio processes given by (45):

$$X_i^{\{l_x\}} = \sum_{j=1}^{l_x} X_i^j,$$

where $X_i^j \in \mathcal{F}_i^S \times \mathcal{F}_i^{M, x, j}$ for $1 \leq j \leq l_x$ and $1 \leq i \leq N$. The aggregate rider value process is

$$U_i^{M, \{l_x\}} = \sum_{j=1}^{l_x} U_i^{M, j} = \mathcal{A}_i^{l_x, x} U_i,$$

since $U_i^j = 0$ if $A_i^{x, j} = 0$. We define two processes $\{X_i^* = E_{\mathbb{P}^M}[X_i^{\{1\}}]\}_{i=0}^N$ and $\{U_i^* = E_{\mathbb{P}^M}[U_i^{M, \{1\}}]\}_{i=0}^N$, both of which lie in $(\Omega^S, \mathcal{F}_N^S, \mathbb{F}^S, \mathbb{Q}^S)$. Then by the SLLN we have

$$\left\{ \frac{X_i^{\{l_x\}}}{l_x} \right\} \rightarrow \{X_i^*\} \quad \text{and} \quad \left\{ \frac{U_i^{M, \{l_x\}}}{l_x} \right\} \rightarrow \{U_i^*\}$$

\mathbb{P}^M -a.s., as $l_x \rightarrow \infty$. Beginning with $X_0^* = 0$, from (46) we have

$$\begin{aligned} X_i^* &= X_{i-1}^* e^{\bar{r}} + {}_{i-1} p_x \left[\Delta_{i-1} (S_i - S_{i-1} e^{\bar{r}}) - C_{i-1} e^{\bar{r}} + p_{x+i-1} \left[F_i - (G - W_{i-} e^{-\bar{\alpha}})^+ \right] \right. \\ &\quad \left. - q_{x+i-1} \left[(DB_{i-1} - W_{i-})^+ \mathbf{1}_{\{\tau \geq i\}} + G \ddot{a}_{\overline{N-i+1}|} \mathbf{1}_{\{\tau < i\}} \right] \right] \end{aligned}$$

for $i \in \mathcal{I}_N^+$. It is immediate that $U_i^* = {}_i p_x U_i$. Finally, from Theorem 17 we have

$$X_i^* = U_i^*$$

\mathbb{Q}^S -a.s., for $i \in \mathcal{I}_N$.

Mortality risk diversification is attained in the limit as $l_x \rightarrow \infty$, and we have perfect hedging. The fair fee was determined assuming optimal surrender behaviour on the part of each policyholder, given survival. If policyholders act irrationally then the insurer can consume from each portfolio at each occurrence of this irrationality. The limiting aggregate portfolio process for the pool is constructed on the basis of homogeneous behaviour of all policyholders, whether or not they act rationally.

Remark 7. *The limiting process was obtained assuming homogeneous policies. This assumption can be weakened to allow for varying initial premiums P by policy, although each policy must have an issue age of x and a common rider fee α . This is true since P can be scaled out of all the processes and the rider fee is independent of the premium P . Let the premium for policy i be P_i . Suppose $\{P_i; i \geq 1\}$ satisfies $\sum_{i=1}^n P_i \rightarrow \infty$ as $n \rightarrow \infty$. Further assume that $\{P_i\}$ is monotonically increasing and satisfies $\sup_{n \geq 1} \frac{n P_n}{\sum_{i=1}^n P_i} < \infty$ or that $\{P_i\}$ is monotonically decreasing in which case no condition is needed. From Theorem 1 in Etemadi [11], as $l_x \rightarrow \infty$, we have*

$$\frac{\sum_{j=1}^{l_x} P_j A_i^{x,j}}{\sum_{j=1}^{l_x} P_j} \rightarrow {}_i p_x$$

\mathbb{P}^M -a.s. for all $i \in \mathcal{I}_N$. Therefore

$$\left\{ \frac{X_i^{\{l_x\}}}{\sum_{i=1}^{l_x} P_i} \right\} \rightarrow \{X_i^*\},$$

with a similar result for U^* . The average is taken on a per premium dollar basis and both X^* and U^* have $P = 1$.

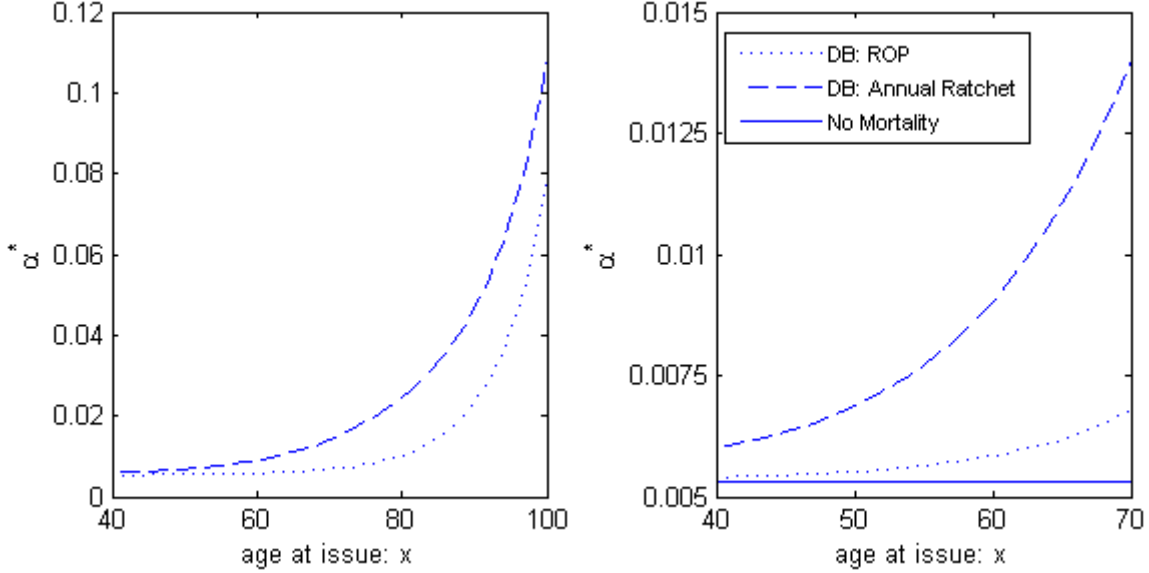
6.5 Numerical Results

We consider two examples. The mortality is modelled using Example 1.

Example 2. Figure 9 plots the fair rider fee α^* against the issue age x for a GMWB with a return of premium DB and an annual ratchet DB without lapses. The parameters are: $g = 7.14\%$, $T = 14$, $r = 5\%$, $\sigma = 20\%$, and $\delta t = 1$. The ratchet adds considerably more value to the contract. The figure on the right zooms in on the ages 40-70. The GMWB plus return of premium DB rider is largely insensitive to x . The payouts upon death or survival are fairly similar in this instance. Under the binomial model without mortality, we have $\alpha^* = 53\text{bps}$ or $V_0(100, 53\text{bps}) = 100$. For the return of premium DB with $x = 60$, we have $\alpha^* = 58\text{bps}$ and $V_0^M(100, 53\text{bps}) = 100.35$. Depending on the product specifications and parameters, mortality may have only a small effect.

Example 3. The diversifiable mortality risk assumption is often quick to be used in the literature. Given the prescribed portfolio process (43) which assumes the risk is diversifiable, we consider the hedging losses when there are only a finite number of policies sold. For $l_x \in \{10, 1000, 100000\}$ we simulated the time of deaths for each policy to obtain $\{\hat{T}_j^x\}_{1 \leq j \leq l_x}$, and computed the average losses per policy per \$100 premium for each path in the binomial model. The parameters used are: $x = 60$, $g = 10\%$, $T = 10$, $r = 5\%$, $\sigma = 15\%$, $\delta t = 1$, and $P = 100$. Surrenders are not allowed.

For the GMWB with an annual ratchet DB, Figure 10 shows the convergence of the hedging losses to zero under the delta hedging strategy as l_x increases. The values are time-zero present values and

Figure 9: α^* as a function of issue age x

the losses under no hedging are also displayed. Figure 11 plots the losses for the limiting portfolio X^* . Table 14 provides the profit metrics $E_{\mathbb{Q}}[\Pi|\{\hat{T}_j^x\}_{1 \leq j \leq l_x}]$ and $SD_{\mathbb{Q}}[\Pi|\{\hat{T}_j^x\}_{1 \leq j \leq l_x}]$ for hedging and no hedging where Π is the average profit per policy discounted to $t = 0$. The results are also given when the DB rider is a return of premium (ROP). The results for both DBs were obtained using the same sets of simulated death times. The column with $l_x = \infty$ represents the results for X^* . The fair fee with the ratchet is 57bps and with the ROP is 44bps. The metrics were calculated using the exact binomial distribution under \mathbb{Q} for the financial risk and the simulated deaths for the mortality risk. For the purpose of examining convergence with respect to l_x , we assume no market price of risk (i.e. $\mu = r$).

Selling a limited number of policies or facing capacity constraints does not impose a significant risk to the insurer in this case because the payouts are similar upon death or survival and diversification occurs rapidly. The average hedging profits are higher with the ROP, but the profits (losses) have more volatility with the ratchet since it pays higher benefits and has higher fees. Under \mathbb{Q} the expected profits are equal under hedging and no hedging. It is the variance that is reduced by hedging.

Without mortality risk, each policy in the pool is subject to a common equity risk and in the binomial world the correct hedging strategy works for any number of policies. Mortality risk introduces incompleteness into the model. Under the assumption of mortality risk diversification the market regains completeness. This occurs in the limit by selling sufficiently large pools of relatively small contract sizes.

Aside from risk pooling and diversification, other risk-management options are reinsurance and longevity bonds. Additionally the typical large life insurer with significant amounts of underwritten business in life insurance and annuities has a degree of natural risk reduction since these instruments have partially offsetting risks. Assuming none of these options are available - there are no re-insurers, longevity bonds do not exist and the insurer only sells annuities - the insurer's main tool for mitigating its risk exposure is by selling a large number of policies of relatively small amounts, thus reducing fluctuations in the realized mortality rates around the expected mortality rates.

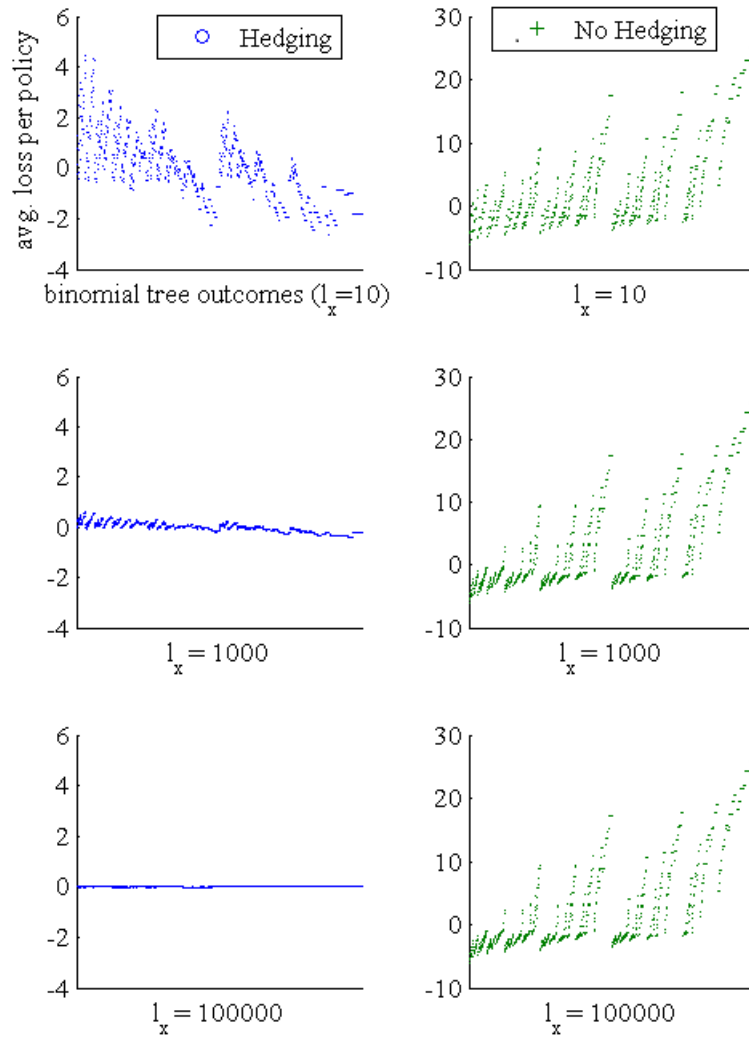
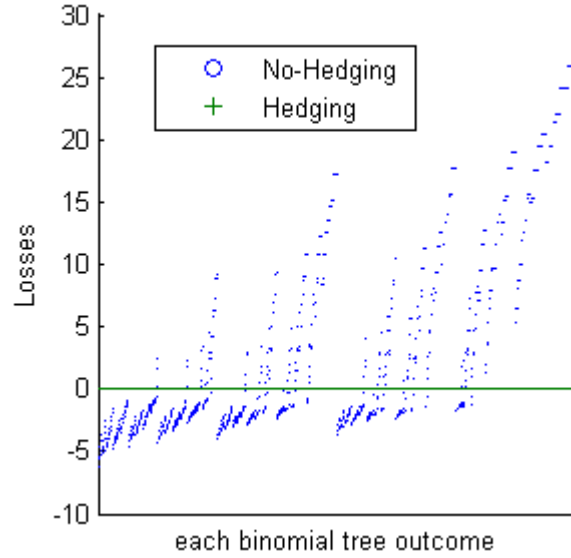


Figure 10: Convergence of losses for GMWB plus ratchet DB as $l_x \rightarrow \infty$ where the average losses per policy under simulated mortality are shown for each market outcome.

Figure 11: Losses for GMWB plus ratchet DB with complete diversification (X^*)

Values per \$100	Hedging			No Hedging			
l_x	10	1000	100000	10	1000	100000	∞
GMWB + Ratchet DB							
$E_{\mathbb{Q}}[\Pi \{\widehat{T}_j^x\}_{1 \leq j \leq l_x}]$	0.122	0.030	0.004	0.122	0.030	0.004	0
$SD_{\mathbb{Q}}[\Pi \{\widehat{T}_j^x\}_{1 \leq j \leq l_x}]$	0.768	0.175	0.008	5.631	5.787	5.860	5.860
GMWB + Return of Premium DB							
$E_{\mathbb{Q}}[\Pi \{\widehat{T}_j^x\}_{1 \leq j \leq l_x}]$	0.261	0.054	0.001	0.261	0.054	0.001	0
$SD_{\mathbb{Q}}[\Pi \{\widehat{T}_j^x\}_{1 \leq j \leq l_x}]$	0.446	0.091	0.004	5.560	5.736	5.776	5.777

Table 14: Profit metrics with and without hedging, with GMDBs

7 Conclusions

In this paper we have constructed a binomial asset pricing model for the variable annuity with GMWB rider which incorporated optimal policyholder surrender behaviour. We extend the continuous time results of Hyndman and Wenger [15] to the discrete-time binomial model by considering the valuation perspectives of the insured and the insurer. These extensions allow us to prove the existence and uniqueness of the fair rider fee and decompose the value of the variable annuity with GMWB rider into term-certain payments and embedded derivatives. Further, in the discrete time binomial model we are able to provide explicit perfect hedging strategies and optimal surrender strategies.

From a computational perspective the ability to model early surrenders using the basic tools of binomial models is one distinct advantage over Monte Carlo methods. The other advantage was demonstrated by easily obtaining an explicit hedging strategy in a binomial (CRR) world that was proved to perfectly hedge the product. A drawback of the binomial model is the $O(2^N)$ growth of the non-recombining binomial trees. Nevertheless, by the tractability of the model and its finite nature, it is straightforward to obtain numerical results concerning any aspect of the product, provided that the number of time-steps is manageable. The qualitative conclusions drawn from such an analysis will usually hold true in the more general continuous model. We present comprehensive numerical results which are consistent with those presented in more complex models.

The binomial modeling framework is further extended to account for diversifiable mortality risk. The diversification argument for mortality risk is sometimes abused in the literature. After applying diversification arguments to obtain the fair fee and hedging results, we imposed capacity constraints by considering finite pools and saw that diversification occurs fairly rapidly. The results support the common claim that insurers are able to diversify mortality risk.

Acknowledgements This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Fonds de recherche du Québec - Nature et technologies (FQRNT).

Appendix A Proofs of Technical Results

Proof of Lemma 4. From the equivalent expression for $v(i, x)$ in (31), the continuity result is immediate. The maximum possible value for $W_N^{x,i}$ is obtained by the path corresponding to $\omega_j = u$ for all $j = i + 1, \dots, N$. Thus

$$b^{x,i} = \min\{\alpha \geq 0 : W_N^{x,i}(uu \dots u) = 0\}.$$

From (30), $W_N^{x,i}(uu \dots u) = 0$ if and only if

$$f(\alpha) := \left(x(e^{-\bar{\alpha}}u)^{N-i} - G \sum_{j=0}^{N-i-1} (e^{-\bar{\alpha}}u)^j \right) \leq 0.$$

But $f \in C^\infty$ and $\lim_{\alpha \rightarrow \infty} f(\alpha) = -G < 0$. We have $f(0) > 0$ if and only if (10) holds. If $f(0) > 0$ then there exists $0 < b^{x,i} < \infty$. If $f(0) \leq 0$, then $b^{x,i} = 0$. The remainder of the proof is similar to the proof of Hyndman and Wenger [15, Lemma 4]. Assume (i, x) is such that $b^{x,i} > 0$. Let

$$A^\alpha := \{W_N^{x,i}(\alpha) > 0\}.$$

Then $A^\alpha \neq \emptyset$ for $\alpha < b^{x,i}$. Fix $\alpha \in [0, b^{x,i})$ and consider $\alpha^{(1)}$ such that $\alpha < \alpha^{(1)} < b^{x,i}$. When restricted to the set $A^{\alpha^{(1)}}$, (30) implies

$$0 < W_N^{x,i}(\alpha^{(1)}) < W_N^{x,i}(\alpha),$$

which in turn implies $A^{\alpha^{(1)}} \subseteq A^\alpha$. We conclude that $v(i, x; \alpha^{(1)}) < v(i, x; \alpha)$. \square

Proof of Theorem 5. From Lemma 4, for $\alpha \geq b^{P,0} > 0$, we have $V_0(P, \alpha, g) = Ga_{\overline{N}} < P$ for $r > 0$. By the definition of U in (13) we have $U \geq 0$ for $\alpha = 0$. By Theorem 7,

$$V_0(P, \alpha = 0, g) = U_0(P, \alpha = 0, g) + P \geq P.$$

By the continuity and strictly decreasing property from Lemma 4, there exists a unique $\alpha^* \in [0, b^{P,0})$. \square

Proof of Theorem 7. We apply backward induction and show that $v(i, x) = u(i, x) + x$ for all $(i, x) \in \mathcal{I}_N \times \mathbb{R}_+$. By definition $v(N, x) = u(N, x) + x$ for all $x \in \mathbb{R}_+$. Assume $v(i, x) = u(i, x) + x$ holds for all $x \in \mathbb{R}_+$ for some $1 \leq i \leq N$. We need to show that $v(i-1, y) = u(i-1, y) + y$ for all $y \in \mathbb{R}_+$. Applying the induction hypothesis,

$$\begin{aligned} v(i-1, y) &= e^{-\bar{r}}[G + pv(i, w(uy)) + qv(i, w(dy))] \\ &= e^{-\bar{r}}[pu(i, w(uy)) + qu(i, w(dy)) + p(w(uy) + G) + q(w(dy) + G)]. \end{aligned}$$

From equations (14) and (15) we have

$$\begin{aligned} u(i-1, y) &= e^{-\bar{r}} \left\{ pu(i, w(uy)) + qu(i, w(dy)) \right. \\ &\quad \left. + p[(G - uye^{-\bar{\alpha}})^+ - uy(1 - e^{-\bar{\alpha}})] + q[(G - dye^{-\bar{\alpha}})^+ - dy(1 - e^{-\bar{\alpha}})] \right\}. \end{aligned}$$

Observe

$$w(y) - (G - ye^{-\bar{\alpha}})^+ = ye^{-\bar{\alpha}} - G.$$

Then

$$w(y) + G - (G - ye^{-\bar{\alpha}})^+ + y(1 - e^{-\bar{\alpha}}) = y,$$

therefore

$$v(i-1, y) - u(i-1, y) = e^{-\bar{r}}[puy + qdy] = y$$

since $pu + qd = e^{\bar{r}}$ by the definition of the risk-neutral probabilities (3). Therefore

$$v(i-1, y) = u(i-1, y) + y$$

for all $y \in \mathbb{R}_+$ and the result holds. \square

Proof of Theorem 11. Following the approach of Shreve [23], we proceed by induction. By assumption we have that $X_0 = U_0$. Assume for some $0 \leq i < N$ that $X_i = U_i$. We need to show that for all $\bar{\omega}_i$,

$$\begin{aligned} X_{i+1}(\bar{\omega}_i u) &= U_{i+1}(\bar{\omega}_i u), \\ X_{i+1}(\bar{\omega}_i d) &= U_{i+1}(\bar{\omega}_i d). \end{aligned}$$

We omit the $\bar{\omega}_i$ notation for conciseness. Substituting U_i for X_i in (29), using (16), (28), and the fact $q = \frac{u - e^{\bar{r}}}{u - d}$ we obtain

$$\begin{aligned} X_{i+1}(u) &= \Delta_i S_i(u - e^{\bar{r}}) + (U_i - C_i)e^{\bar{r}} + F_{i+1}(u) - (G - W_{i+1}(u)e^{-\bar{\alpha}})^+ \\ &= q[u^-(i+1, uW_i) - u^-(i+1, dW_i)] + (pu^-(i+1, uW_i) + qu^-(i+1, dW_i) \\ &\quad + F_{i+1}(u) - (G - W_i u e^{-\bar{\alpha}})^+ \\ &= u^-(i+1, uW_i) + F_{i+1}(u) - (G - W_i u e^{-\bar{\alpha}})^+ \\ &= u(i+1, w(uW_i)) \\ &= U_{i+1}(u). \end{aligned}$$

A similar argument shows that $X_{i+1}(d) = U_{i+1}(d)$. Since $\bar{\omega}_i$ was arbitrary we have $X_{i+1} = U_{i+1}$ and the result holds. \square

Proof of Theorem 17. We proceed by induction. By assumption we have that $X_0 = U_0^M$. Suppose that $E_{\mathbb{P}^M}[X_i] = E_{\mathbb{P}^M}[U_i^M] \mathbb{Q}^S$ -a.s. for some $i \in \mathcal{I}_{N-1}$. For a process H_i we write $H_i(\bar{\omega}_i; j)$ for its value at time i for the specific path $\bar{\omega}_i \omega_{i+1} \dots \omega_N \in \Omega^S$ (where ω_j can take any value in $\{u, d\}$ for all $j > i$) and the specific set $(K^x)^{-1}(j) \in \mathcal{F}_N^{M, \{x, 1\}}$. For any fixed $\bar{\omega}_i$ we need to show that

$$\begin{cases} E_{\mathbb{P}^M}[X_{i+1}(\bar{\omega}_i u; K^x)] = E_{\mathbb{P}^M}[U_{i+1}^M(\bar{\omega}_i u; K^x)], \\ E_{\mathbb{P}^M}[X_{i+1}(\bar{\omega}_i d; K^x)] = E_{\mathbb{P}^M}[U_{i+1}^M(\bar{\omega}_i d; K^x)]. \end{cases}$$

We prove the first equality, the second one is shown in an identical manner. For conciseness, we omit $\bar{\omega}_i$.

Observe that $E_{\mathbb{P}^M}[U_{i+1}^M(u; K^x)] = {}_{i+1}p_x U_{i+1}(u)$. Also $X_{i+1}(u; j) = X_{i+1}(u; K^x > i+1)$ for all $j > i+1$, since $X_{i+1} \in \mathcal{F}_{i+1}$. From (45) we have $X_{i+1}(u; j) = X_i(; j)e^{\bar{r}}$ for all $j \leq i$. Therefore

$$\begin{aligned} E_{\mathbb{P}^M}[X_{i+1}(u, K^x)] &= \sum_{j=1}^N {}_{j-1}q_x X_{i+1}(u, j) + {}_N p_x X_{i+1}(u, K^x > N) \\ &= \sum_{j=1}^i {}_{j-1}q_x X_i(; j)e^{\bar{r}} + {}_i q_x X_{i+1}(u, i+1) + {}_{i+1}p_x X_{i+1}(u, K^x > i+1). \end{aligned}$$

After applying (45) to $X_{i+1}(u, i+1)$ and $X_{i+1}(u; K^x > i+1)$, we obtain

$$\begin{aligned}
E_{\mathbb{P}^M}[X_{i+1}(u, K^x)] &= E_{\mathbb{P}^M}[X_i(; K^x)]e^{\bar{r}} + {}_i p_x [\Delta_i S_i(u - e^{\bar{r}}) - C_i e^{\bar{r}} \\
&\quad - {}_i p_{x+i}((G - W_i u e^{-\bar{\alpha}})^+ - W_i u(1 - e^{-\bar{\alpha}})) \\
&\quad - q_{x+i}((DB_i - W_i u)^+ \mathbf{1}_{\{\tau > i\}} + G\ddot{a}_{\overline{N-i}} \mathbf{1}_{\{\tau \leq i\}})].
\end{aligned} \tag{46}$$

By the induction hypothesis,

$$E_{\mathbb{P}^M}[X_i(; K^x)]e^{\bar{r}} = {}_i p_x U_i e^{\bar{r}}.$$

Then substituting (43) and (38) and applying (39) (in the form U_{i+1}^- , but conditioning on $\tau > i$), we have

$$\begin{aligned}
E_{\mathbb{P}^M}[X_{i+1}(u, K^x)] &= {}_i p_x [(U_i - C_i)e^{\bar{r}} + (U_{i+1}^-(u) - U_{i+1}^-(d))q - Gp_{x+i} \mathbf{1}_{\{\tau \leq i\}} \\
&\quad + \mathbf{1}_{\{\tau > i\}}(p_{x+i}U_{i+1}(u) - U_{i+1}^-(u)) - q_{x+i}(G\ddot{a}_{\overline{N-i}} \mathbf{1}_{\{\tau \leq i\}})] \\
&= {}_i p_x [U_{i+1}^-(u) \mathbf{1}_{\{\tau \leq i\}} - Gp_{x+i} \mathbf{1}_{\{\tau \leq i\}} + \mathbf{1}_{\{\tau > i\}}p_{x+i}U_{i+1}(u) \\
&\quad - q_{x+i}G\ddot{a}_{\overline{N-i}} \mathbf{1}_{\{\tau \leq i\}}] \\
&= {}_{i+1} p_x [\mathbf{1}_{\{\tau > i\}}U_{i+1}(u) + Ga_{\overline{N-(i+1)}} \mathbf{1}_{\{\tau \leq i\}}] \\
&= {}_{i+1} p_x U_{i+1}(u).
\end{aligned}$$

This completes the proof. \square

References

- [1] A. R. Bacinello. Endogenous model of surrender conditions in equity-linked life insurance. *Insurance: Mathematics and Economics*, 37(2):270–296, 2005.
- [2] A. R. Bacinello, P. Millosovich, A. Olivieri, and E. Pitacco. Variable annuities: A unifying valuation approach. *Insurance: Mathematics and Economics*, 49(3):285–297, 2011.
- [3] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–654, 1973.
- [4] N. Bowers, H. Gerber, J. Hickman, D. Jones, and C. Nesbitt. *Actuarial Mathematics*. Society of Actuaries, Schaumburg, Illinois, 2nd edition, 1997.
- [5] P. Boyle and E. S. Schwartz. Equilibrium prices of guarantees under equity-linked contracts. *Journal of Risk and Insurance*, 44(4):639–660, 1977.
- [6] Z. Chen, K. Vetzal, and P. A. Forsyth. The effect of modelling parameters on the value of GMWB guarantees. *Insurance: Mathematics and Economics*, 43(1):165–173, 2008.
- [7] M. Costabile, I. Massabo, and E. Russo. An adjusted binomial model for pricing Asian options. *Review of Quantitative Finance and Accounting*, 27(3):285–296, 2006.

- [8] J. C. Cox, S. A. Ross, and M. Rubinstein. Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263, 1979.
- [9] M. Dai, Y. K. Kwok, and J. Zong. Guaranteed minimum withdrawal benefit in variable annuities. *Mathematical Finance*, 18(4):595–611, 2008.
- [10] D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, NJ, 3rd edition, 2001.
- [11] N. Etemadi. Convergence of weighted averages of random variables revisited. *Proceedings of the American Mathematical Society*, 134(9):2739–2744, 2006.
- [12] R. Geske and K. Shastri. Valuation by approximation: A comparison of alternative option valuation techniques. *The Journal of Financial and Quantitative Analysis*, 20(1):45–71, 1985.
- [13] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York, 2004.
- [14] J. Hull and A. White. Efficient procedures for valuing European and American path-dependent options. *Journal of Derivatives*, 1(1):21–31, 1993.
- [15] C. Hyndman and M. Wenger. Valuation perspectives and decompositions for variable annuities with gmwb riders. *Insurance: Mathematics and Economics*, 55:283–290, March 2014.
- [16] A. Kling, F. Ruez, and J. Ruß. The impact of policyholder behavior on pricing, hedging, and hedge efficiency of withdrawal benefit guarantees in variable annuities. *European Actuarial Journal*, 4(2):281–314, 2014.
- [17] J. Li and A. Szimayer. The effect of policyholders rationality on unit-linked life insurance contracts with surrender guarantees. *Quantitative Finance*, 14(2):327–342, 2014.
- [18] Y. Liu. *Pricing and Hedging the Guaranteed Minimum Withdrawal Benefits in Variable Annuities*. PhD thesis, University of Waterloo, 2010.
- [19] M. A. Milevsky and T. S. Salisbury. Financial valuation of guaranteed minimum withdrawal benefits. *Insurance: Mathematics and Economics*, 38(1):21–38, 2006.
- [20] M. A. Milevsky, S. D. Promislow, and V. R. Young. Killing the law of large numbers: Mortality risk premiums and the Sharpe ratio. *Journal of Risk and Insurance*, 73(4):673–686, 2006.
- [21] T. Møller. Risk-minimizing hedging strategies for unit-linked life insurance contracts. *ASTIN Bulletin*, 28(1):17–47, 1998.
- [22] J. Peng, K. S. Leung, and Y. K. Kwok. Pricing guaranteed minimum withdrawal benefits under stochastic interest rates. *Quantitative Finance*, 12(6):933–941, 2012.
- [23] S. E. Shreve. *Stochastic Calculus for Finance I: The Binomial Asset Pricing Model*. Springer Finance, New York, 2004.
- [24] S. E. Shreve. *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer Finance, New York, 2004.
- [25] S. Wang. *Longevity Risks: Modelling and Financial Engineering*. PhD thesis, Ulm University, 2008.